SLOVENSKÁ ŠTATISTIKA a DEMOGRAFIA

SLOVAK STATISTICS and DEMOGRAPHY

2/2022
ročník/volume 32

Recenzovaný vedecký časopis so zameraním na prezentáciu moderných štatistických a demografických metód a postupov.

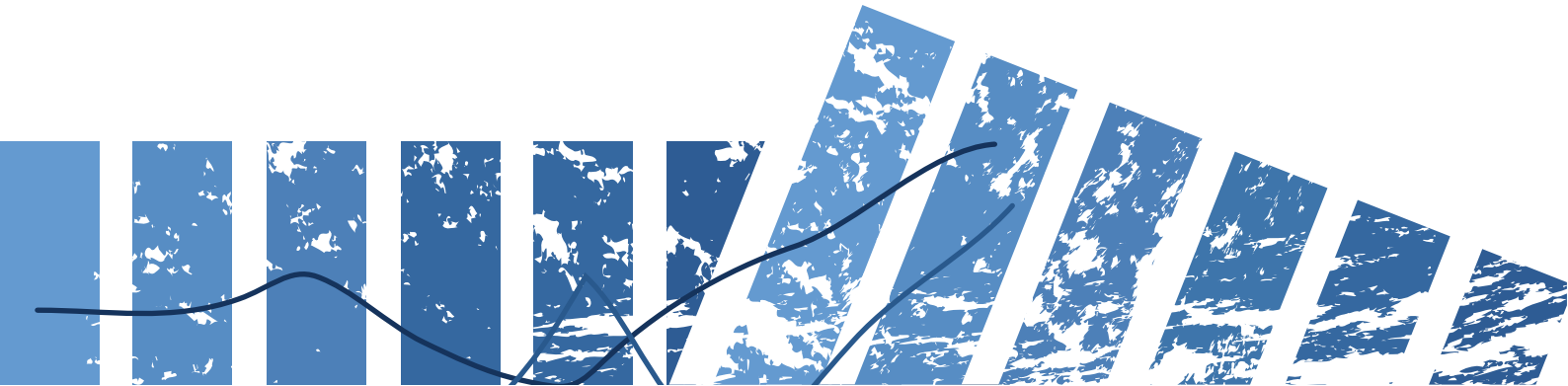
Scientific peer-reviewed journal focusing on the presentation of modern statistical and demographic methods and procedures.

Článok/Article: 3

Typ článku/Type of article: vedecký článok/scientific article

Strany/Pages: 34 – 51

Dátum vydania/Publication date: 15. apríl 2022/April 15, 2022



Milan TEREK
Vysoká škola manažmentu

ANALÝZA KAUZÁLNYCH VZŤAHOV MEDZI PREMENNÝMI

ANALYSIS OF CAUSAL RELATIONS BETWEEN VARIABLES

ABSTRAKT
Overovanie a odhadovanie predpokladaných kauzálnych vzťahov medzi premennými sa považuje za centrálny problém. Cieľom článku je poskytnúť prehľad o existujúcich postupoch a metódach, ktoré možno v tejto oblasti uplatniť. Sú v ňom opísané ich základné črty. Možnosti ich využívania sú ilustrované na jednoduchých príkladoch. To môže uľahčiť proces výberu vhodných metód pri riešení konkrétnych problémov overovania a odhadovania predpokladaných kauzálnych vzťahov medzi premennými.

ABSTRACT
Verifying and estimating the presumed causal relations between the variables is considered a central problem. The aim of the article is to provide an overview of the existing procedures and methods applicable in this area. The article describes their basic features and illustrates the possibilities of their use by simple examples. This can facilitate the process of selecting appropriate methods for solving specific problems of verification and the estimation of the presumed causal relationships between variables.

KĽÚČOVÉ SLOVÁ
kauzálne vzťahy medzi premennými, Simpsonov paradox, analýza ciest, štruktúrne rovnicové modely, kauzálna indukcia.

KEY WORDS
Causal relations between variables, Simpson paradox, path analysis, structural equation models, causal inference.

1. ÚVOD
Príspevok je venovaný možnostiam overovania a odhadovania predpokladaných kauzálnych vzťahov medzi premennými. Analýzu kauzálnych vzťahov medzi premennými možno považovať za centrálny problém. Opíšeme základné črty postupov a metód, ktoré možno v tejto oblasti uplatniť. Možnosti ich aplikácie budeme ilustrovať na jednoduchých príkladoch.

Všimneme si najprv pojmy asociácia a kauzalita. Keď určité hodnoty jednej premennej majú tendenciu objavovať sa spolu s určitými hodnotami druhej premennej, hovoríme že medzi premennými je asociácia alebo spojenie (*association*). Definícia kauzality, ktorú by všeobecne akceptovali štatistici, filozofi aj vedci z iných oblastí, nie je známa. Uvedieme najprv jednu, trochu metaforickú definíciu, ktorá ale poskytuje veľmi dobrú intuitívnu predstavu o kauzalite [10, s. 5]: „Premenná X má kauzálny vplyv na premennú Y, keď Y počúva X a rozhoduje o svojej hodnote na základe toho, čo počuje“. Uvedieme aj presnejšiu definíciu. Zápisom $X \rightarrow Y$ označíme, že premenná X má kauzálny vplyv na premennú Y. Kauzalitu možno potom definovať takto. Keď $X \rightarrow Y$, potom keď sa zmení X, zmení sa rozdelenie Y [3]. Kauzálne vzťahy sú najčastejšie asymetrické. Jedna premenná ovplyvňuje druhú, ale nie opačne.

Samotná asociácia neimplikuje kauzalitu. Kauzálny vzťah medzi premennými implikuje asociáciu, ale opačná implikácia neplatí. Keď je medzi premennými asociácia, môže ale nemusí medzi nimi byť kauzálny vzťah. V [7] je na s. 44 − 45, uvedený zaujímavý príklad. Uvažuje sa o dvoch premenných − počet mentálne postihnutých osôb pripadajúcich na 10 000 obyvateľov v Spojenom kráľovstve a počet vydaných licencií na rádiový prijímač v Spojenom kráľovstve. Sú známe hodnoty týchto premenných v rokoch 1924 až 1937. Skúmal sa lineárny vzťah medzi závisle premennou počet mentálne postihnutých osôb pripadajúcich na 10 000 obyvateľov v Spojenom kráľovstve a nezávisle premennou počet vydaných licencií na rádiový prijímač v Spojenom kráľovstve. Ukázalo sa, že vzťah je štatisticky významný. Koeficient determinácie vyšiel 0,9842. To znamená, že 98,42 % variability závisle premennej možno vysvetliť lineárnym vzťahom medzi počtom mentálne postihnutých osôb pripadajúcich na 10 000 obyvateľov v Spojenom kráľovstve a počtom vydaných licencií na rádiový prijímač v Spojenom kráľovstve. Je zrejmé, že nejde o kauzálny vzťah. Možno jednoducho povedať, že dôvodom štatistického vzťahu medzi týmito premennými je to, že sú spojené monotónnosťou (*monotonically related*).[1]

Všeobecne, vzťah medzi premennými možno považovať za kauzálny, keď sú splnené aspoň tieto tri podmienky [4, s. 288]:

- medzi premennými je asociácia,
- príčina predchádza v čase následok,
- boli vylúčené alternatívne vysvetlenia asociácie.

Prvým predpokladom kauzálneho vzťahu je asociácia medzi premennými. Napríklad vzťah medzi rakovinou pľúc a fajčením v regresných analýzach je veľmi silný. To je však len jeden z predpokladov. Je prirodzené, že predpokladaná príčina musí v kauzálnom vzťahu predchádzať následok. Napríklad ľudia často ochorejú na rakovinu pľúc po rokoch fajčenia. Keď je výskumná štúdia experimentálna[2], možno časovú následnosť v experimente fixovať. V štúdiách založených na pozorovaniach to nie je možné. Vylúčenie alternatívnych vysvetlení asociácie je obyčajne najzložitejší problém. V štúdiách založených na pozorovaniach zvyčajne nie je zložité nájsť asociácie medzi premennými. Tieto asociácie však možno často vysvetliť inými premennými, ktoré v štúdii neboli merané. Napríklad v mnohých medicínskych štúdiách našli asociáciu medzi pitím kávy a pravdepodobnosťou infarktu myokardu. Keď však v štúdiách zohľadnili aj iné premenné, ako napríklad zamestnanie, úroveň stresu a pod., asociácia medzi pitím kávy a pravdepodobnosťou infarktu sa stratila alebo výrazne zmenšila. Preto je splnenie tejto podmienky kauzálneho vzťahu najviac problematické. Môžeme sa nazdávať, že sme vylúčili všetky iné vysvetlenia asociácie, ale istí si nemôžeme byť nikdy. Preto nemôžeme nikdy dokázať, že jedna premenná je v kauzálnom vzťahu s druhou. Hypotézu o kauzalite môžeme len vyvrátiť tým, že ukážeme, že empirický dôkaz je v rozpore aspoň s jedným z troch predpokladov kauzality [3, s. 303].

Okrem troch uvedených podmienok existencie kauzality medzi premennými niektorí autori uvádzajú aj niektoré ďalšie podmienky ktoré podporujú indície o existencii kauzálneho vzťahu medzi premennými. V [9] sa na s. 301-302 uvádza aj podmienka

[1] Dve postupnosti čísiel sú spojené monotónnosťou, keď s rastom jednej postupnosti druhá postupnosť rastie alebo klesá ([7, s. 45]).
[2] O experimentálnej štúdii a štúdii založenej na pozorovaniach pozri v [16, s. 19 - 20].

konzistencie vzťahu medzi premennými. Napríklad veľké množstvo štúdií v rozličných krajinách prezentuje vzťah medzi rakovinou pľúc a fajčením. Charakter vzťahu je v rozličných štúdiách rovnaký. Napríklad vo všetkých štúdiách vychádza, že ľudia, ktorí fajčia viac cigariet denne alebo fajčia dlhší čas, majú rakovinu pľúc častejšie. Ďalšou podmienkou je, aby predpokladaná príčina bola možná. Napríklad experimenty so zvieratami ukazujú, že decht z cigariet spôsobuje rakovinu. Pri splnení všetkých podmienok sú indície o existencii a charaktere kauzálneho vzťahu často veľmi silné, ale obyčajne nie také silné ako na základe výsledkov realizácie dobre navrhnutého experimentu.[3]

Najprv si všimneme možnosti overovania a odhadovania predpokladaných kauzálnych vzťahov medzi jednou závisle premennou a jednou alebo viacerými nezávisle premennými. Tu sa zameriame na základný problém – skúmanie alternatívnych vysvetlení asociácie - a na jeden špecifický problém – Simpsonov paradox. Potom budeme študovať predpokladané kauzálne vzťahy v modeli systému kauzálnych vzťahov, v ktorom môžu byť premenné, o ktorých sa predpokladá, že sú následkom iných premenných a zároveň ovplyvňujú nejaké ďalšie premenné. V tejto súvislosti budeme charakterizovať analýzu ciest a štruktúrne rovnicové modely. Napokon opíšeme základné črty kauzálnej indukcie, ktorá umožňuje nájsť model systému kauzálnych vzťahov, ktorý je v súlade s množinou pozorovaných dát a na ktorom možno realizovať simulačné experimenty a pomocou nich odhadovať účinky zásahov do systému. Problematika analýzy kauzálnych vzťahov je v našej literatúre zatiaľ málo frekventovaná.

2. SKÚMANIE ALTERNATÍVNYCH VYSVETLENÍ ASOCIÁCIE POMOCOU KONTROLY INÝCH PREMENNÝCH

Predpokladajme, že premenná *X* má kauzálny vplyv na premennú *Y*. Základným komponentom overovania platnosti tohto predpokladu je skúmanie možných alternatívnych vysvetlení asociácie. Možno ho realizovať prostredníctvom skúmania, či sa asociácia medzi *X* a *Y* nestratí, keď odstránime účinky iných premenných na túto asociáciu. Premennú, ktorej vplyv odstránime vo viacrozmernej analýze obyčajne nazývame kontrolovaná (*controlled*) alebo kontrolná (*control*) premenná. Na rozdiel od technických a prírodných vied sa v spoločenskovednej oblasti vo výskume väčšinou využívajú štúdie založené na pozorovaniach. Nemožno jednoducho fixovať hodnoty premenných, ktoré chceme kontrolovať, ako je to bežné v experimentálnych štúdiách. Možno však zoskupiť hodnoty pozorovaní do skupín s rovnakými alebo podobnými hodnotami kontrolovaných premenných. Ak je kontrolovanou premennou napríklad vzdelanie, môžeme rozdeliť pozorovania povedzme do troch skupín: pozorovania na jednotkách s nižším ako stredoškolským vzdelaním, pozorovania na jednotkách so stredoškolským vzdelaním a pozorovania na jednotkách s vysokoškolským vzdelaním. To je štatistická kontrola. V [3] sa na s. 304 - 305 uvádza príklad štúdia asociácie medzi výškou študenta a počtom bodov z matematického testu na základe náhodného výberu študentov jednej americkej strednej školy. Koeficient korelácie vyšiel 0,81, ide teda o silný priamy vzťah, s rastom výšky študenta má počet bodov z testu tendenciu rásť. Môže byť výška študenta príčinou lepších výsledkov z matematiky? Hľadal sa spôsob, ako inak možno túto asociáciu vysvetliť. V náhodnom výbere boli študenti rozličného veku. Starší študenti majú tendenciu byť vyšší a mať aj hlbšie znalosti z matematiky. Účinok veku na asociáciu možno odstrániť prostredníctvom štatistickej

[3] *O navrhovaní experimentov pozri [8, s. 563 – 645] a [16, s. 319 – 346].*

kontroly tak, že analyzujeme asociáciu medzi výškou študenta a počtom bodov z matematického testu zvlášť pre každú vekovú kategóriu študentov. Ukázalo sa, že korelácia medzi výškou študenta a počtom bodov z matematického testu pre študentov z rovnakej vekovej kategórie bola blízka nule pre každú z troch uvažovaných vekových kategórií. To je empirický dôkaz toho, že výška študenta nemá kauzálny účinok na výsledok matematického testu, pretože asociácia zmizla, keď bola veková kategória študentov konštantná. Všeobecne sa študuje vzťah medzi *X* a *Y* pre prípady s rovnakou alebo podobnou hodnotou kontrolnej premennej. Tým, že hodnota kontrolnej premennej je konštantná, odstráni sa vplyv tejto premennej na asociáciu medzi *X* a *Y*.

Premenná, ktorá sa v štúdii nemeria, ale ovplyvňuje skúmanú asociáciu, sa niekedy nazýva skrytá premenná (*lurking variable*). Možno súčasne uvažovať o viacerých vysvetľujúcich a kontrolných premenných, ktoré označíme X_1, X_2, ... Hovoríme, že asociácia medzi X_1 a *Y* je zdanlivá (*spurious*), keď obe premenné závisia od tretej premennej X_2. Zdanlivá asociácia nie je jediná možnosť pri ktorej asociácia zmizne keď kontrolujeme tretiu premennú. Ďalšou možnosťou je kauzálny reťazec (*chain of causation*) v ktorom X_1 ovplyvňuje X_2, ktoré zasa ovplyvňuje *Y.* Premenná X_2 sa niekedy nazýva sprostredkovacia premenná (*intervening* alebo *mediator variable*). Niekedy dve premenné nevykazujú asociáciu, pokiaľ sa nekontroluje iná premenná. Táto kontrolná premenná sa najčastejšie nazýva odrušovacia (*suppressor variable*). Často sa účinok vysvetľujúcej premennej na *Y* mení podľa úrovne inej vysvetľujúcej alebo kontrolnej premennej. Keď sa účinok premennej X_1 na *Y* mení na rôznych úrovniach X_2, hovoríme že X_1 a X_2 sú v interakcii. Vtedy má asociácia rozličnú silu a /alebo charakter[4] na rozličných úrovniach X_2. Keď dve vysvetľujúce premenné ovplyvňujú *Y* , ale sú aj navzájom spojené, hovoríme o zmiešaní účinkov (*confounding*). Vtedy je ťažké určiť, ktorá z nich v skutočnosti ovplyvňuje *Y,* pretože účinok jednej z nich môže byť čiastočne daný jej asociáciou s druhou premennou. Obyčajne pozorujeme iný účinok na *Y* pre jednu premennú, keď kontrolujeme druhú premennú, ako keď ju nekontrolujeme. Zmiešanie účinkov je v spoločenskovednom výskume celkom bežné. Preto je napríklad ťažké identifikovať príčiny zlepšovania ekonomických výsledkov a pod. (podrobnejšie o typoch vzťahov vo viacrozmernej analýze, pozri v [17, s. 109 - 111]).

Metódy viacrozmernej analýzy ako napríklad viacnásobná regresia umožňujú vykonať štatistickú kontrolu a ohodnotiť vzory asociácie a interakcie bez vykonania čiastkových analýz na rozličných kombináciách úrovní kontrolných premenných. V minulosti sa predpokladalo, že každý regresný model by mal modelovať kauzálne vzťahy medzi premennými. Regresia, ktorá evidentne nebola modelom kauzálnych vzťahov medzi premennými, sa nazývala zdanlivou. Je však užitočné oddeliť aj v regresii asociáciu od kauzality. Asociáciu medzi premennými v regresii možno využiť, aj keď nedáva jednoznačnú odpoveď na otázku o existencii kauzálnych vzťahov medzi nimi. Totiž odhadnutú regresnú rovnicu možno využiť na odhadovanie a na tvorbu predikcií vtedy, keď sú splnené predpoklady regresnej analýzy, vzťah medzi premennými je štatisticky významný a koeficient determinácie ukazuje, že regresia je kvalitná[5] bez ohľadu na to, či odráža alebo neodráža kauzálne vzťahy

[4] Zmena priameho vzťahu na nepriamy alebo opačne.
[5] Podrobnejšie v [16, s. 243].

medzi premennými do tej miery, do akej vzťahy medzi dátami zistené pri odhadovaní zostávajú nezmenené. To sa široko využíva napríklad pri aplikáciách regresnej analýzy v hĺbkovej analýze dát (*data mining*)[6]. Ak je cieľom nájdenie najlepšieho modelu na predikcie hodnôt nejakej premennej, predikčný model sa formuluje len z dát, bez ohľadu na to, či regresný model odráža, alebo neodráža kauzálne vzťahy medzi premennými. Regresná analýza môže poskytnúť najlepšiu predstavu o existencii a charaktere kauzálnych vzťahov medzi premennými na základe výsledkov realizácie dobre navrhnutého experimentu.[7]

2.1 SIMPSONOV PARADOX

Všimnime si teraz jeden špecifický problém, ktorý komplikuje odhadovanie kauzálnych vzťahov. Je možné, že keď kontrolujeme jednu premennú, každá asociácia v čiastkovej kontingenčnej tabuľke má opačný charakter (podrobnejšie o analýze kategoriálnych dát pozri v [1], [2]). To sa nazýva Simpsonov paradox.

Príklad 1. Máme náhodný výber 800 pacientov, ktorí majú prístup k lieku. 400 pacientov sa rozhodlo liek užívať, 400 pacientov liek neužívalo. Výsledky liečby sú v tabuľke 1. V každom políčku je zľava počet pacientov, v zátvorke je počet uzdravených pacientov a nasleduje percento uzdravených pacientov.

Tabuľka č. 1: Dáta o liečbe

Užívanie lieku Pohlavie	Užíval liek (U)	Neužíval liek (NU)
Muži (M)	112 (106), 94,6 %	295 (253), 85,8 %
Ženy (Z)	288 (213), 74 %	105 (74), 70,5 %
Spolu	400 (319), 79,8 %	400 (327), 81,8 %

Zdroj údajov: vlastné výpočty

Podmienenú pravdepodobnosť, že sa muž vylieči (V), keď užíva liek P(V|U) možno odhadnúť pomerom 106/112 = 0,946. Potom pravdepodobnosť, že sa muž nevylieči (NV), keď užíva liek $P(NV|U) = 1 - P(V|U) = 1 - 0,946 = 0,054$. Podobne P(V|NU) = 0,858 a P(NV|NU) = 0,142. Odhadnutá šanca, že sa muž vylieči keď užíva liek, je 0,946/0,054 = 17,52 a odhadnutá šanca že sa vylieči keď neužíva liek, je 0,858/0,142 = 6,04. Pomer šancí je 17,52/6,04 = 2,9. Muž má približne 2,9 krát väčšiu šancu, že sa vylieči, keď užíva liek, ako keď ho neužíva. Podobne by sme pre ženy vypočítali, že žena má 1,19 krát väčšiu šancu, že sa vylieči, keď užíva liek, ako keď ho neužíva. Keď však uvažujeme o všetkých pacientoch spolu, nedelíme ich na mužov a ženy, vyjde že pacient má 1,14krát väčšiu šancu, že sa uzdraví, keď liek neužíva, ako keď ho užíva. Zdá sa, že keď poznáme pohlavie pacienta, mali by sme liek odporúčať, keď ho nepoznáme, nemali by sme ho odporúčať. Takýto záver je samozrejme nezmyselný. To, že nepoznáme pohlavie pacienta, nemôže z užitočného lieku urobiť neužitočný.

E. Simpson [14] ako prvý publikoval poznatok o existencii dát, v ktorých charakter asociácie v celom súbore je opačný v každom pod-súbore. Podľa neho sa takáto vlastnosť dát začala nazývať Simpsonov paradox. Aby sme mohli rozhodnúť, či liek pacientovi uškodí alebo pomôže, musíme najprv porozumieť tomu, čo stojí za dátami – príčinnému mechanizmu, ktorý generoval výsledky, ktoré vidíme. Predpokladajme,

[6] Podrobnejšie o hĺbkovej analýze dát pozri v [15].
[7] O navrhovaní experimentov v regresii pozri v [16, s. 280 - 281].

že je známe, že estrogén má negatívny vplyv na liečbu, takže je menšia pravdepodobnosť vyliečenia ženy ako muža bez ohľadu na užívanie lieku. Pritom vo výbere pacientov ktorí liek užívali je len 112/400 = 0,28, t. j. 28 % mužov a 72 % žien. Byť ženou možno považovať za spoločnú príčinu užívania lieku aj jeho menšej účinnosti. Aby sme mohli ohodnotiť účinnosť lieku, musíme porovnávať pacientov rovnakého pohlavia, prípadne štruktúra pacientov podľa pohlavia a užívania lieku musí byť rovnaká. Uzdravenie pacienta závisí od pohlavia aj od užívania lieku. Keď vylúčime vplyv pohlavia, teda pohlavie bude kontrolnou premennou, dostaneme asociáciu medzi uzdravením pacienta a užívaním lieku. Liek by sme mali odporúčať.

Príklad 2. V príklade 1 predpokladajme, že namiesto pohlavia pacientov sa zaznamenával tlak krvi. Predpokladajme, že je známe, že liek ovplyvňuje vyliečenie prostredníctvom zníženia krvného tlaku u tých, ktorí ho užívajú - ale žiaľ, má aj toxický účinok. Dostali sme výsledky uvedené v tabuľke 2.

Tabuľka č. 2: Dáta o liečbe 1

Užívanie lieku Krvný tlak	Neužíval liek (NU)	Užíval liek (U)
Nízky krvný tlak (NT)	112 (106), 94,6 %	295 (253), 85,8 %
Vysoký krvný tlak (VT)	288 (213), 74 %	105 (74), 70,5 %
Spolu	400 (319), 79,8 %	400 (327), 81,8 %

Zdroj údajov: vlastné výpočty

V tabuľke 2 sú rovnaké dáta ako v tabuľke 1 s tým rozdielom, že oproti tabuľke 1 sú vymenené kategórie stĺpcov a v riadkoch sú namiesto kategórií pohlavia, kategórie krvného tlaku. Mali by sme v tomto prípade liek odporúčať? Odpoveď opäť vychádza zo spôsobu, akým boli dáta vygenerované. V základnom súbore liek môže zlepšiť mieru vyliečenia kvôli svojmu účinku na krvný tlak, v pod súboroch - skupina ľudí, ktorí mali po liečbe nízky krvný tlak a skupina ľudí, ktorí mali po liečbe vysoký krvný tlak, by sme tento účinok nevideli, videli by sme len toxický účinok lieku. Rovnako ako v príklade 1, cieľom experimentu bolo zistiť celkový účinok užívania lieku na mieru vyliečenia. Tu však zníženie krvného tlaku je jedným z faktorov, ktorými liečebná procedúra ovplyvňuje mieru vyliečenia, a preto nemá zmysel triediť výsledky na základe krvného tlaku. Takže v tomto prípade, vychádzajúc z doplnkových informácií by sme sa mali na rozdiel od príkladu 1 oprieť o výsledky získané za celý základný súbor a liek odporúčať. Dáta v príklade 1 a 2 sú rovnaké, správne rozhodnutie v príklade 1 sa opiera o separované dáta, v druhom príklade o súhrnné dáta. Rozdiel je v doplnkových informáciách, ktoré opisujú príčinný mechanizmus, ktorý generuje dáta. Bez jeho znalosti, len na základe štatistickej analýzy dát by sme nevedeli prijať správne rozhodnutie. Keď skúmame asociáciu medzi premennými s cieľom dozvedieť sa niečo o vzťahoch medzi premennými, nakoniec vždy vychádzame z nejakej základnej predstavy o kauzálnych súvislostiach, ktorá vychádza z nejakých doplnkových informácií.

V experimentálnych štúdiách možno experiment naplánovať tak, aby sme sa vyhli Simpsonovmu paradoxu, generovanému zmiešaním účinkov, samozrejme opäť len na základe vhodných doplnkových informácií (pozri napr. [5]).

3. MODELOVANIE SYSTÉMU KAUZÁLNYCH VZŤAHOV

Opíšeme niektoré možnosti modelovania systému kauzálnych vzťahov a ukážeme, aké predpoklady sú nevyhnutné na odhadovanie ich parametrov v štúdiách založených na pozorovaniach.

3.1 ANALÝZA CIEST

Analýza ciest (*path analysis*) používa na vyjadrenie predpokladov o príčinných vzťahoch medzi premennými, regresné modely, ktoré zahŕňajú príslušné kontrolné premenné. Analytik musí explicitne špecifikovať predpoklady o kauzálnych vzťahoch medzi premennými [4, s. 506]. Začiatočná predstava o kauzálnych vzťahoch medzi premennými často obsahuje taký systém vzťahov, v ktorých vystupujú premenné o ktorých sa predpokladá, že sú dôsledkom iných premenných, ale zároveň ovplyvňujú nejaké ďalšie premenné. Jeden model viacnásobnej regresie nedokáže opísať takýto systém vzťahov, pretože obsahuje len jednu závisle premennú. Analýza ciest využíva sústavu regresných modelov, ktorá dokáže zachytiť všetky predpokladané vzťahy medzi premennými.

Základným nástrojom na opísanie začiatočnej predstavy o kauzálnych vzťahoch medzi premennými je diagram ciest (*path diagram*). V diagrame ciest je predpokladaný kauzálny vzťah vyjadrený orientovanou hranou, ktorá vychádza z uzla, reprezentujúceho príčinu (nezávisle premenná) a končí sa v uzle, reprezentujúcom následok (závisle premenná). Závisle premenné v príslušných regresných rovniciach korešpondujú s uzlami, v ktorých sa končí aspoň jedna orientovaná hrana. Diagram ciest vyjadrí hypotézy o tom, ktoré premenné sú „zodpovedné“ za asociáciu dvoch premenných.

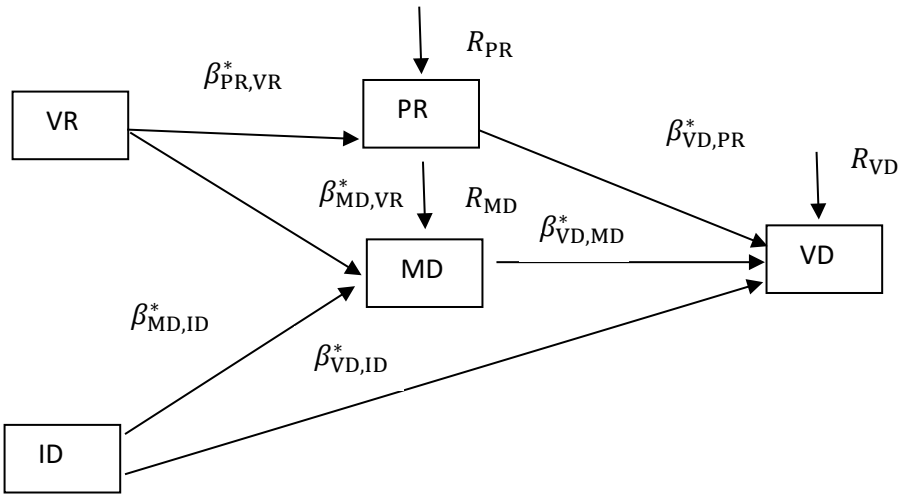
V [4, s. 506 - 509] sa uvádza príklad o analýze vzťahov medzi dosiahnutým vzdelaním dieťaťa VD, jeho motiváciou MD, inteligenciou ID, vzdelaním rodičov VR a príjmami rodičov PR. Predpokladajú sa kauzálne väzby, zobrazené na obrázku 1. Uzly reprezentujú premenné, orientované hrany reprezentujú predpokladané kauzálne väzby. Koeficienty nad hranami (*path coefficients*) sú normované regresné koeficienty (*standardized regression coefficients*). Normovaný regresný koeficient β_i^* pri nezávisle premennej X_i udáva, o koľko svojich smerodajných odchýlok sa zmení stredná hodnota závisle premennej *Y*, keď sa premenná X_i zväčší o jednu smerodajnú odchýlku *Y*, keď sú všetky ostatné nezávisle premenné konštantné (kontrolované). Tieto koeficienty sú vhodné na porovnanie relatívnych účinkov viacerých nezávisle premenných na závisle premennú, pretože sa merajú v rovnakých merných jednotkách (v smerodajných odchýlkach závisle premennej *Y*). Diagram ciest na obrázku č. 1 korešponduje so sústavou troch regresných rovníc:

$$E(\mathrm{VD}) = \beta^*_{\mathrm{VD,PR}} \cdot \mathrm{PR} + \beta^*_{\mathrm{VD,MD}} \cdot \mathrm{MD} + \beta^*_{\mathrm{VD,ID}} \cdot \mathrm{ID}$$

$$E(\mathrm{MD}) = \beta^*_{\mathrm{MD,VR}} \cdot \mathrm{VR} + \beta^*_{\mathrm{MD,ID}} \cdot \mathrm{ID}$$

$$E(\mathrm{PR}) = \beta^*_{\mathrm{PR,VR}} \cdot \mathrm{VR}$$

Obrázok č. 1 Diagram ciest dosiahnutého vzdelania dieťaťa



Zdroj: *[4,* ***s. 506 – 509], upravené***

Každá zo závisle premenných je v diagrame ciest spojená so svojou reziduálnou premennou (*residual path variable*) [13]. Tieto premenné reprezentujú variabilitu, nevysvetlenú nezávisle premennými v rovnici. Každá z reziduálnych premenných reprezentuje zostávajúci podiel (1 – R^2) nevysvetlenej variability závisle premennej, kde R^2 je koeficient determinácie pre regresnú rovnicu závisle premennej. Normovaný regresný koeficient tejto premennej sa rovná $\sqrt{1-R^2}$ [4, s. 507]. Jedným zo základných výsledkov analýzy ciest je dekompozícia korelácie medzi dvomi premennými na časti, ktoré súvisia s rozličnými cestami medzi týmito dvomi premennými (podrobnejšie v [4, s. 509 – 510]). Analýza ciest sa všeobecne realizuje v troch krokoch:

1. Predbežná predstava o kauzálnych vzťahoch sa vyjadrí pomocou diagramu ciest, bez normovaných regresných koeficientov.
2. Vytvorí sa sústava regresných modelov ktorá slúži na odhadnutie normovaných regresných koeficientov premenných, vrátane reziduálnych premenných.
3. Prostredníctvom kontroly, či sú dáta z náhodného výberu v súlade s modelom, sa model ohodnotí. Prípadné nevýznamné cesty sa zrušia. Modifikovaný model slúži na ďalšiu analýzu. Odhadnú sa normované regresné koeficienty pre tento modifikovaný model.

Na korektné použitie vzťahu na dekompozíciu ciest je v každom z regresných modelov sústavy nevyhnutné splniť predpoklad, že nemerané premenné, ktoré sú reprezentované reziduálnou premennou pre závisle premennú v regresnom modeli sú nekorelované s nezávisle premennými v tomto modeli. V príklade by napríklad všetky nemerané premenné, ktoré ovplyvňujú dosiahnuté vzdelanie dieťaťa mali byť nekorelované s príjmom rodičov, motiváciou dieťaťa a inteligenciou dieťaťa. To je príliš „silná“ požiadavka. Všeobecne je splnenie tohto predpokladu málo pravdepodobné a prakticky neoveriteľné. Tiež je pravdepodobné, že keby sa do regresných modelov pridali ďalšie premenné, získali by sme odlišný obraz o možných kauzálnych vzťahoch. Nakoniec, ak sú dáta z náhodného výberu aj v súlade s formulovaným diagramom ciest, nedokazuje to pravdivosť kauzálneho systému reprezentovaného týmto diagramom ciest. Pomocou štatistických metód nemožno priamo testovať

predpokladané kauzálne väzby. Ani analýza ciest neumožňuje z asociácie indukovať kauzalitu. Poskytuje len štruktúru na reprezentáciu a odhadovanie predpokladaných kauzálnych účinkov [4, s. 510].

3.2 ŠTRUKTÚRNE ROVNICOVÉ MODELY

Všeobecný štruktúrny rovnicový model (*Structural equation model*) kombinuje prvky analýzy ciest a faktorovej analýzy (podrobnejšie o faktorovej analýze pozri [6]). Model sa nazýva model štruktúry kovariancie (*covariance structure model*), pretože sa snaží vysvetliť rozptyly a korelácie medzi pozorovanými premennými. Toto vysvetlenie má formu kauzálneho modelu, spájajúceho systém faktorov, z ktorých niektoré môžu byť vytvorené vo faktorovej analýze a niektoré môžu byť pozorované premenné. Modely štruktúry kovariancie majú dva komponenty. Prvý je model merania **(***measurement model***)**. Podobá sa na faktorovú analýzu tým, že vytvára množinu nepozorovaných faktorov z pozorovaných premenných. Druhý komponent je štruktúrny rovnicový model. Ten sa podobá na analýzu ciest tým, že špecifikuje regresné modely pre faktory, vytvorené v modeli merania.

Model merania špecifikuje, ako sú pozorované premenné spojené s množinou latentných premenných[8]. Táto časť analýzy sa podobá na faktorovú analýzu ibaže, modelovanie má viac špecifikovanú štruktúru. Model merania priradí každú latentnú premennú a priori k špecifikovanej množine pozorovaných premenných. To sa dosiahne tak, že saturácie[9] niektorých faktorov sú rovné nule, teda predpokladá sa, že niektoré latentné premenné sú nekorelované s inými premennými. Model merania zohľadňuje fakt, že pozorované premenné môžu byť zaťažené chybami merania a problémami platnosti a spoľahlivosti, a preto nie sú najlepšími ukazovateľmi konceptov, ktoré nás zaujímajú. Účelom vytvorenia latentných premenných je nájdenie operatívnejších charakteristík, namiesto tých, ktoré je ťažké dobre zmerať, ako sú predsudky, úzkosť, konzervativizmus a podobne.

Štruktúrny rovnicový model používa regresné modely na špecifikáciu kauzálnych vzťahov medzi latentnými premennými. Jedna alebo viaceré z latentných premenných sa špecifikujú ako nezávisle premenné. Latentné závisle premenné môžu v regresnom modeli závisieť od latentných nezávisle premenných ako aj od iných latentných závisle premenných. Na rozdiel od analýzy ciest umožňuje tento prístup aproximáciu modelov s dvojcestnou kauzalitou, v ktorých môžu latentné premenné v regresnom modeli závisieť od ľubovoľných iných latentných premenných.

3.3 KAUZÁLNA INDUKCIA

Na pochopenie metód a modelov kauzálnej indukcie (*causal inference*) je nevyhnutná znalosť základov teórie pravdepodobnosti, štatistiky a teórie grafov. Štruktúrne kauzálne modely (*structural causal models* - SCM) slúžia na opis čŕt prírody a toho, ako sú spojené s inými črtami, alebo ako príroda priraďuje hodnoty premenným, ktoré nás zaujímajú. Formálne je štruktúrny kauzálny model zložený z dvoch množín premenných *U* a *V* a z množiny funkcií *f*, ktoré priraďujú každej premennej z *V* hodnotu, v závislosti od hodnôt iných premenných v modeli. Premenné v množine *U* sa nazývajú exogénne premenné (*exogenous variables*), čo znamená, že vstupujú do modelu zvonka, nezaujíma nás, čo je ich príčinou. Premenné v množine

[8] V štatistike sa nepozorované premenné, ako napríklad faktory, obyčajne nazývajú latentné premenné.
[9] Korelácia premennej s faktorom sa nazýva saturácia (loading) premennej týmto faktorom.

V sú endogénne premenné (*endogenous variables*). Každá endogénna premenná je v orientovanom grafe nasledovníkom (*descendant*) aspoň jednej exogénnej premennej. Exogénne premenné nemôžu byť nasledovníkom žiadnych iných premenných. Nemajú v orientovanom grafe predchodcov (*ancestor*) a sú v ňom reprezentované koreňovými uzlami (*root nodes*). Keď je známa hodnota každej exogénnej premennej, možno pomocou funkcií *f* determinovať hodnotu každej endogénnej premennej [10, s. 27], [11].

Každý SCM je spojený s grafickým kauzálnym modelom (*graphical causal model*). Ten pozostáva z množiny uzlov, ktoré reprezentujú premenné v množinách *U* a *V* a z množiny orientovaných hrán, ktoré ich spájajú. Orientované hrany reprezentujú funkcie z množiny *f*. Grafický kauzálny model *G* pre nejaký SCM obsahuje jeden uzol pre každú premennú. Ak funkcia f_X pre premennú *X* (definuje hodnoty premennej *X*) obsahuje aj premennú *Y* (hodnoty *X* závisia aj od *Y*), potom v *G* vedie z *Y* do *X* orientovaná hrana. Uvažuje sa len o orientovaných acyklických grafoch *G*[10].

Keď v grafickom kauzálnom modeli, premenná *Y* je priamym nasledovníkom premennej *X*, potom *X* je priamou príčinou *Y*. Keď *Y* je nasledovníkom *X* (nie priamym), potom *X* je potenciálnou príčinou *Y*. Definíciu kauzality možno spresniť takto: Premenná *X* je priamou príčinou (*direct cause*) premennej *Y*, keď *X* je vo funkcii, ktorá definuje hodnoty premennej *Y*. Premenná *X* je príčinou (*cause*) premennej *Y*, keď je priamou alebo potenciálnou príčinou *Y*.

Môže vzniknúť otázka, na čo vlastne používať grafický model, keď obsahuje menej informácií ako SCM. Po prvé, predstavy o kauzálnych vzťahoch medzi premennými obyčajne nie sú kvantitatívne, ako to vyžaduje SCM, ale kvalitatívne, vyjadriteľné v grafickom modeli. Vieme napríklad, že pohlavie človeka je príčinou jeho výšky (muži sú všeobecne vyšší ako ženy) a že výška človeka je jednou z príčin jeho výkonnosti v basketbale, ale numerické ohodnotenie týchto vzťahov nie je jednoduché. Graf kauzálnych vzťahov poskytne čiastočne špecifikovanú verziu SCM. Výhodou grafických kauzálnych modelov je aj to, že umožňujú veľmi efektívne vyjadriť združené rozdelenia pravdepodobnosti. Priameho nasledovníka v modeli nazveme potomkom a priameho predchodcu nazveme rodičom. Pre ľubovoľný acyklický model, združené rozdelenie pravdepodobnosti premenných v modeli je dané súčinom podmienených rozdelení *P*(potomok | rodičia). Toto pravidlo – pravidlo rozkladu súčinom (*product decomposition rule*) je:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} P(x_i|pa_i) \qquad (1)$$

kde pa_i sú hodnoty rodičov premennej X_i [10, s.29].

Proces, ktorý možno použiť v grafickom kauzálnom modeli akejkoľvek zložitosti s cieľom predpovedať závislosti, ktoré zdieľajú všetky súbory dát generované týmto grafom sa nazýva *d*-separácia (*d-separation*) (písmeno *d* znamená *directional* – smerová). Tento proces umožňuje pri ľubovoľnom páre uzlov určiť, či sú *d*-spojené (*d-connected*), to znamená, či medzi nimi existuje cesta, alebo sú *d*-separované (*d-separated*), čo znamená, že medzi nimi neexistuje cesta. Keď povieme, že dva uzly sú *d*-separované, myslíme tým, že premenné, ktoré reprezentujú, sú určite nezávislé.

[10] Graf, ktorý ako podgraf neobsahuje kružnicu (uzavretú postupnosť prepojených vrcholov), sa nazýva acyklický.

Keď povieme, že dva uzly sú *d*-spojené, myslíme tým, že premenné, ktoré tieto uzly reprezentujú sú možno alebo s najväčšou pravdepodobnosťou závislé [10, s. 46].[11]

3.3.1 TESTOVANIE MODELU A HĽADANIE KAUZÁLNYCH VZŤAHOV

Keď máme graf *G* o ktorom sa domnievame, že generuje množinu dát *S*, *d*-separácia ukáže, ktoré premenné musia byť podmienene nezávislé[12], v závislosti od ktorých iných premenných. Podmienenú nezávislosť možno testovať pomocou dát. Predpokladajme, že máme zoznam podmienok *d*-separácie v *G* a vidíme, že premenné *A* a *B* musia byť podmienene nezávislé, v závislosti od *C* (*independent conditional on C*). Predpokladajme, že odhadneme pravdepodobnosti na základe *S* a zistíme, že dáta naznačujú, že *A* a *B* nie sú podmienene nezávislé v závislosti od *C*. Potom môžeme zamietnuť *G* ako možný kauzálny model pre *S* [10, s. 48 – 49].

Oproti iným metódam testovania adekvátnosti modelu má *d*-separácia niekoľko výhod. Je neparametrická – nevyžaduje špecifické funkcie na spájanie premenných. Testuje model lokálne, nie globálne. Umožňuje špecifikovať spojenia, v ktorých model nie je v súlade s dátami z výberu a naznačuje, ako možno model opraviť. S využitím *d*-separácie možno mnoho modelov vylúčiť a dospieť k množine modelov, ktoré nie sú v rozpore s pozorovanými dátami. Význam tohto výsledku spočíva v tom, že dovoľuje nájsť pre kauzálne modely množinu dát, ktorú by mohli generovať. Nielenže možno začať s kauzálnym modelom a generovať množinu dát, ale možno začať s množinou dát a odôvodniť ich pôvod kauzálnym modelom. To je veľmi užitočné, pretože cieľom väčšiny výskumov zameraných na dáta je nájsť model, ktorý ich vysvetľuje.

3.3.2 ÚČINKY ZÁSAHOV

Konečným cieľom mnohých štatistických štúdií je predpovedať účinky zásahov. Keď napríklad zhromažďujeme dáta spojené s povodňami, v konečnom dôsledku nám ide o nájdenie spôsobov, ako možno znížiť ich škodlivé dôsledky a frekvenciu. Tu sa možno oprieť len o dáta zo štúdií založených na pozorovaniach, nie o výsledky realizácie experimentov. Keď kontrolujeme premennú v modeli, fixujeme jej hodnotu. Meníme systém a výsledkom je často zmena hodnôt ďalších premenných. Keď premennú podmienime, nič nezmeníme; iba zúžime náš pohľad na podmnožinu prípadov, v ktorých premenná má hodnotu, ktorá nás zaujíma. To, čo sa potom mení, je naše vnímanie sveta, nie svet samotný. V kauzálnej indukcii sa rozlišuje medzi prípadom, keď premenná *X* nadobudne hodnotu *x* prirodzene ($X = x$), a prípadom, keď sa premenná fixuje na hodnote *x*, zápisom $do(X = x)$. Takže $P(Y = y \mid X = x)$ je pravdepodobnosť že $Y = y$, keď sme zistili, že $X = x$, zatiaľ čo $P(Y = y \mid do(X = x))$ je pravdepodobnosť, že $Y = y$, keď premennú *X* fixujeme na hodnote *x*. Pravdepodobnosť $P(Y = y \mid X = x)$ je pravdepodobnosť rozdelenia pravdepodobnosti *Y* jednotiek, ktoré majú hodnotu *x* premennej *X*. Pravdepodobnosť $P(Y = y \mid do(X = x))$ je pravdepodobnosť rozdelenia pravdepodobnosti *Y* jednotiek, keď hodnotu *X* všetkých jednotiek v základnom súbore fixujeme na úrovni *x*. Mlčky sa predpokladá, že intervencia nemá žiadne „vedľajšie účinky“, to znamená že priradenie hodnoty *x* premennej *X* pre nejakú jednotku nemení priamym spôsobom premenné, ktoré

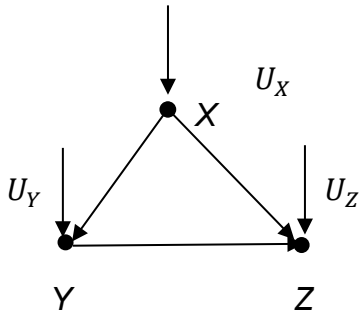
[11] Definuje sa aj pojem pravdepodobná závislosť: Premenné Z a Y sú pravdepodobne závislé (likely dependent), keď pre niektoré z, y: $P(Z = z|Y = y) \neq P(Z = z)$.

[12] Podmienená závislosť a nezávislosť premenných je definovaná rozlične v rozličných konfiguráciách premenných- reťazce (chains), vidlice (forks), kolízne schémy (colliders).

nasledujú. Napríklad dobrovoľné užívanie lieku môže mať iný vplyv na liečbu jednotlivca ako vynútené užívanie, v rozpore s náboženskými zásadami človeka. Ak sú vedľajšie účinky prítomné, musia byť explicitne zahrnuté do modelu.

Príklad 1 – pokračovanie 1. Máme grafický model, ktorý reprezentuje problém z príkladu 1. Premenná *X* je pohlavie, *Y* je užívanie lieku a *Z* je uzdravenie.

Obrázok č. 2: Grafický kauzálny model, ktorý reprezentuje účinok nového liečiva



Zdroj: [10, s. 55], upravené

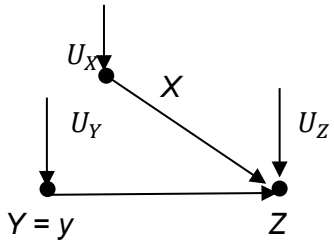
Premenné U_X, U_Y, U_Z na obrázku č. 2, sú exogénne premenné – chybové členy. Ak chceme zistiť, aký účinný je liek v celom základnom súbore, predstavme si hypotetický zásah, ktorým liek podávame všetkým jednotkám základného súboru. Výsledok porovnáme s mierou vyliečenia, keď sa liek nepodáva. Označme prvý zásah ako *do* (*Y* = 1) a druhý ako *do* (*Y* = 0). Úlohou je odhadnúť rozdiel:

$$P(Z = 1|do(Y = 1)) - P(Z = 1|do(Y = 0))$$

ktorý je známy ako rozdiel v kauzálnom účinku (*causal effect difference*), alebo priemerný kauzálny účinok (*average causal effect* - ACE).

Je známe, že príčinné účinky nemožno odhadnúť zo samotného súboru dát, bez „kauzálneho príbehu“. Samotné dáta z pozorovacej štúdie nestačia na zistenie, či je účinok lieku pozitívny alebo negatívny. Pomocou grafu na obrázku 2 však možno vypočítať veľkosť kauzálneho účinku z dát. Aby sme to dosiahli, simulujeme zásah na obrázku 3. Kauzálny účinok P ($Z = z$ | *do* ($Y = y$)) sa rovná podmienenej pravdepodobnosti $P_m(Z = z|Y = y)$ z manipulovaného (*manipulated*) modelu na obrázku č. 3.

Obrázok č. 3: Modifikovaný grafický model, ktorý predstavuje zásah do modelu na obrázku 2



Zdroj: [10, s. 56], upravené

Model na obrázku 3 nastavuje užívanie lieku v základnom súbore. Výsledkom je manipulovaná pravdepodobnosť P_m. Kľúč k výpočtu príčinného účinku spočíva v tom, že manipulovaná pravdepodobnosť P_m má dve rovnaké základné vlastnosti s pôvodnou pravdepodobnostnou funkciou v modeli na obrázku 2. Po prvé, marginálna pravdepodobnosť $P(X = x)$ je invariantná k intervencii, pretože proces ktorý determinuje X, nie je ovplyvnený odstránením hrany z X do Y. V príklade to znamená, že proporcie mužov a žien pred intervenciou a po nej zostávajú rovnaké. Po druhé, podmienená pravdepodobnosť $P(Z = z|X = x, Y = y)$ je invariantná, pretože proces, prostredníctvom ktorého Z reaguje na X a Y, $Z = f(y, x, u_z)$, zostáva rovnaký bez ohľadu na to, či sa Y mení spontánne alebo v dôsledku zámernej manipulácie. Môžeme teda napísať dve rovnice invariancie:

$$P_m(Z = z|X = x, Y = y) = P(Z = z|X = x, Y = y) \quad \text{a} \quad P_m(X = x) = P(X = x).$$

Možno využiť aj skutočnosť, že X a Y sú v modifikovanom modeli d-separované, a teda nezávislé. To hovorí, že

$$P_m(X = x|Y = y) = P_m(X = x) = P(X = x).$$

Keď tieto úvahy spojíme, dostaneme:

$$\begin{gathered} P(Z = z|do(Y = y) = \\ = P_m(Z = z|Y = y) = \\ = \sum_{\boldsymbol{x}} P_m(Z = z|Y = y, X = x) P_m(X = x|Y = y) = \\ = \sum_{\boldsymbol{x}} P_m(Z = z|Y = y, X = x) P_m(X = x) \end{gathered}$$

Nakoniec, keď použijeme vzťahy invariancie, dostaneme vzťah na výpočet kauzálneho účinku v termínoch predintervenčných pravdepodobností:

$$P(Z = z|do(Y = y)) = \sum_x P(Z = z|Y = y, X = x) P(X = x) \qquad (2)$$

Rovnica (2) sa nazýva vzorec úpravy (*adjustment formula*). Možno podľa neho vypočítať silu asociácie medzi Y a Z pre každú hodnotu x premennej X ($X = 1$ – muž, $X = 0$ – žena), ktorá sa potom spriemeruje cez tieto hodnoty. Tento postup sa označuje ako úprava pre X (*adjusting for X*) alebo kontrola pre X (*controlling for X*).

Pravú stranu v rovnici (2) možno odhadnúť priamo z dát. Pre $Y = 1$ – pacient užíval liek, $Z = 1$ – pacient sa uzdravil, podľa (2) dostaneme

$$P(Z = 1|do(Y = 1)) = P(Z = 1|Y = 1, X = 1)P(X = 1) + P(Z = 1|Y = 1, X = 0)P(X = 0).$$

Po dosadení do predchádzajúceho vzťahu z tabuľky 1, dostaneme

$$P(Z = 1|do(Y = 1)) = 0{,}946\ (112 + 295)/800 + 0{,}740(288 + 105)/800 = 0{,}4812775 + 0{,}363525 = 0{,}8448025$$

Podobne, pre $Y = 0$ – pacient neužíval liek, $Z = 1$ – pacient sa uzdravil, podľa (2) dostaneme

$$P(Z = 1|do(Y = 0)) = P(Z = 1|Y = 0, X = 1)P(X = 1) + P(Z = 1|Y = 0, X = 0)P(X = 0)$$

Po dosadení do predchádzajúceho vzťahu z tabuľky 1, dostaneme

$P(Z = 1|do(Y = 0)) = 0{,}858(112 + 295)/800 + 0{,}705(288 + 105)/800 = 0{,}4365075 +$
$+ 0{,}34633125 = 0{,}78283875$

Keď porovnáme účinok užívania lieku ($Y = 1$) s účinkom jeho neužívania ($Y = 0$), dostaneme

$\text{ACE} = P(Z = 1|do(Y = 1)) - P(Z = 1|do(Y = 0)) = 0{,}8448025 - 0{,}78283875 = ´$
$= 0{,}06196375$

Z výsledku je zrejmý pozitívny vplyv užívania lieku. Užívanie lieku zväčší pravdepodobnosť uzdravenia v priemere o 0,06196375.

4. ZÁVER

Najsilnejšie indície o existencii a charaktere kauzálnych vzťahov medzi premennými možno získať na základe výsledkov realizácie dobre navrhnutého experimentu, v ktorom možno merať účinok zmeny vstupných premenných na výstupnú premennú. V sociálnoekonomickej oblasti však väčšinou nie je možné navrhnúť a realizovať experimentálne štúdie, pretože niektoré premenné ktoré ovplyvňujú výstupnú premennú nie je možné kontrolovať. Vtedy sa možno oprieť len o štúdie založené na pozorovaniach. Problém týchto štúdií spočíva v tom, že je ťažké oddeliť koreláciu od príčinných súvislostí. Kauzálny vzťah medzi premennými implikuje asociáciu medzi nimi, opačná implikácia však neplatí. Keď sa opierame o dáta zo štúdií založených na pozorovaniach, potom vzťah medzi premennými možno považovať za kauzálny, keď je medzi premennými asociácia, príčina predchádza v čase následok a boli vylúčené alternatívne vysvetlenia asociácie. Splnenie poslednej podmienky je najviac problematické. Nikdy si totiž nemôžeme byť istí, že sme vylúčili všetky iné možné vysvetlenia asociácie a preto nemôžeme nikdy dokázať, že jedna premenná je v kauzálnom vzťahu s druhou.

Pri hľadaní indícií o existencii a charaktere kauzálnych vzťahov medzi premennými je nevyhnutné oprieť sa o nejaké doplnkové informácie umožňujúce opísať predpokladaný príčinný mechanizmus, ktorý generuje dáta. Len na základe štatistickej analýzy dát zo štúdie založenej na pozorovaniach nemožno takéto indície získať. Za každým príčinným záverom musí byť nejaký príčinný predpoklad, ktorý nie je testovateľný na základe dát zo štúdií založených na pozorovaniach [12, s. 99]. V príklade o asociácii medzi počtom mentálne postihnutých osôb pripadajúcich na 10 000 obyvateľov a počtom vydaných licencií na rádiový prijímač v Spojenom kráľovstve sme nakoniec postupovali podobne. Jednoducho sme implicitne konštatovali, že nepoznáme žiadny príčinný mechanizmus, ktorý by mohol generovať dané dáta, a teda napriek asociácii medzi premennými, nejde o kauzálny vzťah.

Ak chceme študovať predpokladané kauzálne vzťahy v modeli systému kauzálnych vzťahov medzi premennými, možno využiť analýzu ciest, ktorá vyjadruje predpoklady o kauzálnych vzťahoch medzi premennými pomocou sústavy regresných modelov. Štruktúrne rovnicové modely kombinujú prvky analýzy ciest a faktorovej analýzy, pričom majú formu kauzálneho modelu ktorý sa týka systému faktorov, z ktorých niektoré mohli byť vytvorené vo faktorovej analýze a niektoré môžu byť pozorované premenné. Podobne ako v prípade analýzy ciest, ani pomocou štruktúrnych

rovnicových modelov nemožno priamo testovať predpokladané kauzálne väzby. Aj keď ide o sofistikovanejšiu metódu, ani ona nedokáže indukovať kauzalitu z asociácie. Poskytuje, podobne ako analýza ciest, len štruktúru na reprezentáciu a odhadovanie predpokladaných kauzálnych účinkov.

Za najkomplexnejšiu metodológiu na analýzu kauzálnych vzťahov treba určite považovať kauzálnu indukciu, ktorá umožňuje základnú predstavu o kauzálnych vzťahoch medzi premennými vyjadriť pomocou štruktúrneho kauzálneho modelu, ktorý je spojený s grafickým kauzálnym modelom. V grafickom kauzálnom modeli možno pomocou *d*-separácie špecifikovať spojenia, v ktorých model nie je v súlade s dátami z výberu a naznačuje, ako možno model opraviť. Niektoré modely možno vylúčiť a dospieť k množine modelov, ktoré nie sú v rozpore s pozorovanými dátami. Možno začať aj s množinou dát a odôvodniť ich pôvod kauzálnym modelom. To môže byť často veľmi užitočné. Mimoriadnym prínosom tejto metodológie je možnosť realizovať s modelom simulačné experimenty a odhadovať účinky zásahov do systému. Aplikácia tejto metodológie v praxi však vyžaduje veľmi dobrú znalosť teórie pravdepodobnosti, induktívnych štatistických metód, regresnej analýzy a teórie grafov, spolu s niektorými novými poznatkami, ktoré vznikli v rámci nej. Ani táto metodológia však nedokáže poskytnúť dôkazy o existencii a charaktere kauzálnych vzťahov medzi premennými. Umožňuje len nájsť model kauzálnych vzťahov, ktorý nie je v rozpore s pozorovanými dátami a ponúka nástroje na experimentovanie s týmto modelom a odhadovanie účinkov simulovaných zásahov do systému.

Táto práca bola podporená vedeckou grantovou agentúrou KEGA, v rámci projektu K-20-035-00: Learn Economics: aplikácia e-vzdelávania ako novej formy výučby ekonómie.

LITERATÚRA

[1] AGRESTI, A.: Analysis of Ordinal Categorical Data. Second Edition. Hoboken: Wiley and Sons, 2010. ISBN 978-0-470-08289-8.
[2] AGRESTI, A.: Categorical Data Analysis. Third Edition. Hoboken: Wiley and Sons, 2013. ISBN 978-0-470-46363-5.
[3] AGRESTI, A. – FINLAY, B.: Statistical Methods for the Social Sciences. Fourth edition. Essex: Pearson, 2014. ISBN 978-1-29202-166-9.
[4] AGRESTI, A.: Statistical Methods for the Social Sciences. Fifth edition. Boston: Pearson, 2018. ISBN 13: 978-0-13-450710-1.
[5] AMERINGER, S. – SERLIN, R. C. – WARD, S.: Simpson's Paradox and Experimental Research. Dostupné na: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2880329/pdf/nihms-204397.pdf
[6] HÄARDLE, W. – HLÁVKA, Z.: Multivariate Statistics: Exercises and Solutions. 2nd edition. Berlin Heidelberg: Springer-Verlag, 2015. ISBN: 978-3-642-36004-6.
[7] MONTGOMERY, D. C. – PECK, E. A. – VINING, G. G.: Introduction to Linear Regression Analysis. Fifth Edition. Hoboken: J. Wiley and Sons, 2012. ISBN 978-0-470-54281-1.
[8] MONTGOMERY, D. C.: Statistical Quality Control. A Modern Introduction. Seventh edition. Singapore: J. Wiley and Sons, 2013. ISBN 978-1-118-32257-4.
[9] MOORE, D. S. – NOTZ, W. I.: Statistics. Concepts and Controversies. New York: W. H. Freeman and Company, 2006. ISBN 978-0-7167-8636-8.
[10] PEARL, J. – GLYMOUR, M. – JEWELL, N. P.: Causal Inference in Statistics. Chichester: J. Wiley and Sons, 2016. ISBN 9781119186847.

[11] PEARL, J.: Causality: Models, Reasoning, and Inference. 2nd ed. New York: Cambridge University Press, 2009 (1).
[12] PEARL, J.: Causal inference in statistics: An overview. In: Statistics Surveys, Vol. 3, 2009 (2), s. 96–146.
[13] RETHERFORD, R. D. – MINJA KIM CHOE: Statistical Models for Causal Analysis. New York: Wiley and Sons, 1993. ISBN 0-471-55802-8.
[14] SIMPSON, E.: The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, 1951, Series B, 13, s. 238 – 241.
[15] TEREK, M. – HORNÍKOVÁ, A. – LABUDOVÁ, V.: Hĺbková analýza údajov. Bratislava: Iura Edition, 2010. ISBN 978-80-8078-336-5.
[16] TEREK, M.: Interpretácia štatistiky a dát. Piate, doplnené vydanie. Košice: Equilibria, 2017. ISBN 978-80-8143-213-2.
[17] TEREK, M.: Dotazníkové prieskumy a analýzy získaných dát. 1. vydanie. Košice: Equilibria 2019. ISBN 978-80-8143-247-7.

RESUMÉ

Overovanie a odhadovanie predpokladaných kauzálnych vzťahov medzi premennými sa považuje za centrálny problém. Cieľom článku je poskytnúť prehľad o existujúcich postupoch a metódach, ktoré možno v tejto oblasti uplatniť. Sú v ňom opísané ich základné črty. Možnosti ich využívania sú ilustrované na jednoduchých príkladoch.
V článku sú definované pojmy asociácia a kauzalita. Vzťah medzi premennými možno považovať za kauzálny, keď je medzi premennými asociácia, príčina predchádza v čase následok a boli vylúčené alternatívne vysvetlenia asociácie.
Ak predpokladáme, že premenná *X* má kauzálny vplyv na premennú *Y*, základným komponentom overovania platnosti tohto predpokladu je skúmanie možných alternatívnych vysvetlení asociácie. Tie možno skúmať pomocou kontroly iných premenných. V štúdiách založených na pozorovaniach možno zoskupiť hodnoty pozorovaní do skupín s rovnakými alebo podobnými hodnotami kontrolovaných premenných – vykonať štatistickú kontrolu. Metódy viacrozmernej analýzy ako napríklad viacnásobná regresia, umožňujú vykonať štatistickú kontrolu a ohodnotiť vzory asociácie a interakcie bez vykonania čiastkových analýz na rozličných kombináciách úrovní kontrolných premenných. Nikdy nemožno dokázať, že jedna premenná je v kauzálnom vzťahu s druhou. Hypotézu o kauzalite možno len vyvrátiť tým, že empirický dôkaz je v rozpore aspoň s jedným z predpokladov kauzality.
Je možné, že keď kontrolujeme jednu premennú, každá asociácia v čiastkovej kontingenčnej tabuľke má opačný charakter. To sa nazýva Simpsonov paradox. V článku sa uvádzajú dva príklady Simpsonovho paradoxu. Dáta v príkladoch sú rovnaké, ale doplnkové informácie opisujúce príčinný mechanizmus, ktorý generuje dáta sú rozdielne, preto sa v rozhodovaní v prvom príklade opierame o separované dáta, v druhom príklade o súhrnné dáta. Bez znalosti príčinného mechanizmu, len na základe štatistickej analýzy dát sa nevieme správne rozhodnúť.
Pri štúdiu predpokladaných kauzálnych vzťahov v modeli systému kauzálnych vzťahov, v ktorom môžu byť premenné, o ktorých sa predpokladá, že sú dôsledkom iných premenných a zároveň ovplyvňujú nejaké ďalšie premenné, možno využiť analýzu ciest, ktorá vyjadruje predpoklady o kauzálnych vzťahoch medzi premennými pomocou sústavy regresných modelov. Štruktúrne rovnicové modely kombinujú prvky analýzy ciest a faktorovej analýzy. Podobne ako v prípade analýzy ciest, ani pomocou štruktúrnych rovnicových modelov nemožno priamo testovať predpokladané kauzálne väzby. Aj keď ide o sofistikovanejšiu metódu, ani ona nedokáže indukovať kauzalitu,

z asociácie. Poskytuje, podobne ako analýza ciest, len štruktúru na reprezentáciu a odhadovanie predpokladaných kauzálnych účinkov.
Za najkomplexnejšiu metodológiu na analýzu kauzálnych vzťahov možno považovať kauzálnu indukciu, ktorá umožňuje základnú predstavu o kauzálnych vzťahoch medzi premennými vyjadriť pomocou štruktúrneho kauzálneho modelu, spojeného s grafickým kauzálnym modelom. V grafickom kauzálnom modeli možno špecifikovať spojenia, v ktorých model nie je v súlade s dátami z výberu a naznačuje, ako možno model opraviť. Mimoriadnym prínosom tejto metodológie je možnosť realizovať s modelom simulačné experimenty a odhadovať účinky zásahov do systému. Ani táto metodológia však nedokáže poskytnúť dôkazy o existencii a charaktere kauzálnych vzťahov medzi premennými. Umožňuje len nájsť model kauzálnych vzťahov, ktorý nie je v rozpore s pozorovanými dátami a ponúka nástroje na experimentovanie s týmto modelom a odhadovanie účinkov simulovaných zásahov do systému.

RESUME
The author considers the verification and estimation of the presumed causal relations between the variables to be a central problem. The aim of the article is to provide an overview of the existing procedures and methods applicable in this field. The article describes their basic features and illustrates the possibilities of their use by simple examples.
The terms association and causality are defined. The relationship between the variables can be considered causal when there is an association between them, the cause precedes the time consequence and alternative explanations of the association were excluded.
Assuming that the variable X has a causal effect on the variable Y, the basic component of validating this assumption is to examine the possible alternative explanations for the association. These can be examined by controlling other variables. In observation-based studies, the observation values can be divided into groups with the same or similar values of the controlled variables - statistical control can be performed. Multivariate analysis methods such as multiple regression enable to perform statistical control and evaluate patterns of association and interaction without performing partial analyzes on various combinations of control variable levels. It can never be proven that one variable is causally related to another. The hypothesis of causality can only be disproved by the fact that empirical evidence contradicts at least one of the causality assumptions.
It is possible that when we check one variable, each association in the partial contingency table has the opposite character. This is called the Simpson's paradox. Two examples of Simson's paradox are provided in the article. The data in the examples are identical, but the additional information describing the causal mechanism that generates the data is different, so in the first example we rely on separate data, in the second example on the aggregated data. Without knowing the causal mechanism, based only on the statistical data analysis the right decision cannot be made.
In the study of the presumed causal relationships in the model of the system of causal relationships, which may contain variables that are supposed to affect other variables and at the same time also affect some other variables, a path analysis can be used, expressing assumptions about causal relationships between variables using a set of regression models. Structural equation models combine elements of path analysis and factor analysis. Similarly as in case of path analysis, the presumed causal links cannot be directly tested using structural equation models. Although it is a more sophisticated

method, not even that could induce causality from the association. It provides, like path analysis, only a structure for representing and estimating the expected causal effects. Causal inference can be considered the most complex methodology for the analysis of causal relationships, which enables the basic idea of causal relationships between variables to be expressed using a structural causal model, which is associated with a graphical causal model. In the graphical causal model, connections can be specified in which the model does not comply with the sample data and indicates how the model can be corrected. A special benefit of this methodology is the ability to perform simulation experiments with the model and estimate the effects of interventions in the system. However, even this methodology fails to provide proof of the existence and the nature of causal relationships between variables. It only allows to find a model of causal relationships that does not contradict the observed data and offers tools for experimenting with this model and estimating the effects of simulated interventions in the system.

PROFESIJNÝ ŽIVOTOPIS

***Prof. Ing. Milan Terek, PhD.**, od roku 2018 pracuje ako profesor na Vysokej škole manažmentu v Bratislave. Vedie kurzy Úvod do štatistiky, Štatistika, Matematika pre manažérov II, Kvantitatívne metódy pre manažérov a Kvantitatívne metódy vo výskume v oblasti podnikového manažmentu. V rokoch 1977 – 2018 pracoval na Ekonomickej univerzite v Bratislave. Viedol kurzy Štatistika, Štatistické riadenie kvality, Analýza rozhodovania, Hĺbková analýza dát, Výberové skúmanie, Lineárne programovanie, Nelineárne programovanie, Operačný výskum a Systémové modelovanie. Vo výskume sa zameriava na aplikácie štatistických metód v ekonómii a manažmente. Je autorom alebo spoluautorom 6 monografií, 10 vysokoškolských učebníc, 17 skrípt, 75 článkov vo vedeckých a odborných časopisoch a 115 príspevkov na vedeckých konferenciách, publikovaných v zborníkoch.*

KONTAKT

mterek@vsm.sk