Informační Bulletin

České Statistické Společnosti číslo 1., leden 1995, ročník 6.

Povaha věcí se ráda skrývá.
(*Hérakleitos*)

Princip zúplnění

Jan Coufal

V poslední době se velice rozvíjí smysl pro rovnoprávnost, rovnováhu, symetrii, proporce, ale také pro úplnost, důslednost a komplexnost pohledu na svět, jakož i pro odhalování a chápání protikladů[1]. Tím se hluboce a nezvratně utvrzujeme v přesvědčení, že nic si nezasluhuje, aby bylo preferováno na úkor něčeho jiného, že potlačované naopak zasluhuje, aby bylo povzneseno a hýčkáno, a především – neobjevené objeveno, nalezeno, odhaleno.

Na přírodě je možné pozorovat, že svět je mnohotvářný[2], protože tam, kde je vidět jen jedna tvář, je nejen žádoucí, ale i nezbytně nutné, a samozřejmě také možné odhalit druhou, třetí,..., n-tou,... tvář, i kdyby byla pouze odvrácená či zahalená (ale ví se o ní), nebo by byla alespoň tušena, či dokonce o její existenci nebylo ani potuchy. Tzn. je třeba ji vymyslet nebo stvořit. Inu, řekne-li se A, musí se říci nejen B, ale také C, $Č$, D, $Ď$,..., $Ž$. Když se např. v první třídě vyučuje prvouka, musí se také ve druhé třídě vyučovat druhouka, ve třetí třetiuka, ...

To vše lze shrnout v **principu zúplnění**, který platí v celém vesmíru (včetně mikrosvěta) a jeho ε-okolí, kde ε je jakkoli velké.

[1] aniž by se omezoval na dvojice protikladů, protože protikladné je cokoli s čímkoli

[2] někdo uvádí mnohotvárný, ale jak uvidíme dále, jde o chybný termín

Princip zúplnění. *Neexistují věci samy o sobě*[3]*, ale vše existuje vždy samo s něčím*[4]*.*

Demostrace. Na základě zkušeností uveďme. Pokud si necháváme přinést jídlo v restauraci, zpravidla pravíme se vší exaktností: *"Přineste mi něco s něčím."*

Význam. Jak se ukazuje, je význam tohoto principu při vší skromnosti obrovský. Závratnými možnostmi, které nejen nabízí, ale i přímo vnucuje, se stává přínosným, inspirujícím, podnětným, neobyčejně plodným a kreativním ve všech oblastech lidského (a nejen lidského) počínání, zvláště ve všech vědních, technických i uměleckých oborech. Pro ilustraci uveďme několik příkladů jeho aplikace.

Filosofie – zde lze tvůrčím způsobem přispět k řešení kardinální otázky – *co bylo dříve vejce nebo slepice?* Užitím principu zúplnění lze upozornit na klíčovou roli (dosud zcela opomíjenou), kterou v této věci hraje kohout.

Politická ekonomie – kromě práce nutné umožňuje princip zúplnění definovat i práci postačující.

Biologie – jsou-li bílkoviny, musejí býti žloutkoviny; vedle měkkýšů musejí existovat tvrdýši, vedle slepýšů hluchýši; k prvokům nutně patří druhoci, třeťoci, ..., n-ťoci, ...; vedle tetřeva-hlušce existuje tetřev-slepec i tetřev-němec; jednoduše lze definovat velblouda n-hrbého ($n \in \mathbb{N}$).

Anatomie – díky principu lze hravě učinit objev, že slepé střevo je nejen slepé, ale také hluché a němé.

Technika – pomocí principu zúplnění lze sestrojit závodní automobil *Beta-Julie*, zkonstruovat secí stroj na kokosové ořechy i uplatnit metodu leteckého jednocení cukrové řepy atd.

Byrokracie – již v minulosti byl princip zúplnění uplatňován při komplexním hodnocení pracovníků, neboť toto hodnocení mělo reálnou a imaginární část, přičemž imaginární část byla daleko důležitější než reálná.

Vojenství – vedle polního maršála umožňuje princip zavést hodnost lučního maršála, protože (jak známo) na louce lze válčit ještě lépe než na poli.

Urbanismus – např. v Praze lze doplnit čtvrti Servác, Bonifác, Ctirad, Naddědek, Běhov, Rozmařilov atd.

[3] *Ding an sich* je blbost

[4] tj. *Ding an Ding*

Lyžování – vedle běžek zcela jistě existují chodky, stejně většina lyžařů na lyžích chodí.

Kybernetika – dnešní značné úsilí a velká pozornost, které se věnují vytváření umělé inteligence, je nutné naprosto nezbytně rozšířit o zkoumání umělé blbosti. Je faktem, že při zkoumání této se samozřejmě naráží na blbost přirozenou. Argumenty, že vzhledem k obecnému dostatku přirozené blbosti není nutná umělá, nemohou obstát, protože umělá inteligence se také nemůže stát plnohodnotnou náhradou lidských schopností. Zde má právě tato nová disciplína daleko větší šanci na úspěch. Navíc její užití k modelování skutečných objektů je zcela nepochybné. Řešení je ale spojeno s velkými obtížemi, protože ve srovnání s problémy umělé inteligence, které jsou finitní[5], jde o problémy infinitní[6], protože *omezenost je neomezená*.

Školství – nejen život si žádá Fakultu veterinárního práva[7]; je potřebná řečtina na Lesnických fakultách, aby její absolventi mohli vedle myslivecké latiny pěstovat i mysliveckou řečtinu; také je nezbytné studovat obory romská urbanistika, jezdecké umění Aztéků, současná sumerská literatura, antická filatelie, fonetika němého filmu, dějiny antarktického zemědělství, Parmenidova dynamika, Herakleitova statika, dějiny inovačních tradic, tautologická dialektika, ústavy lidových oligarchií, Booleova heuristika, dějiny malířství na Velikonočních ostrovech atd.

Fyzika vesmíru – vesmír se od velkého třesku rozpíná (už asi 20 miliard let). Soudí se, že se za nějakou dobu bude smršťovat. Co se stane, až se vesmír splácne dohromady? Princip zúplnění odpovídá, že nutně nastane velký plesk, po němž bude následovat velký třesk, ... ad infinitum. Tím lze charakterisovat vývoj vesmíru jak **třesky-plesky**.

Umění – je zvláště zarážející s jak přímo trestuhodnou nedbalostí (právě ve vztahu k principu zúplnění) se velmi často setkáváme u jinak důkladných původců ať uměleckých děl či vědeckých objevů. Alespoň namátkou uveďme:

- "•" v Darwinově bibliografii chybí objevné dílo *"O původu družek"*;
- "•" Stendhal nenapsal román *"Modrý, žlutý, zelený, fialový a bílý"*;

[5] tj. konečné

[6] tj. nekonečné

[7] a tím také zavedení titulu *doktor veterinárního práva*, zkratka JVDr.

- "•" v české literatuře chybí horor *Dědeček*, jehož děj by se odehrával na Novém Černidle;
- "•" Lehár nesložil operetu *"Smutný vdovec"*;
- "•" A co Smetana? Kde jsou opery *"Koupený ženich"*, *"Češi v Braniborech"*, *"Nedalibučina"*, *"Dva vdovci"*, *"Kopanec"*, *"Andělův strop"* či *"Mikulášova podlaha"*, *"Přísné tajemství"*, *"Přísné tajemství zvláštní důležitosti"*, *"Housle"*, *"Violoncello"*, *"Basa"*? Kde jsou symfonické básně *Pankrác*, *Labe*, *Podbaba*, *Sezimovo Ústí*, *Macocha* či *Z moravských luhů a hájů*?
- "•" Napsal Vejvoda valčík *Tatra lásky*?
- "•" Leonardo da Vinci nenamaloval obraz *Sterea Lisa*;
- "•" vždyť např. i William Shakespeare nedůsledně napsal slavný Hamletův monolog: *"... Být či nebýt? – toť otázka"*, aniž by si uvědomil, že jde vlastně o stereolog, protože *Být? či Nebýt? – toť dvě otázky*;
- "•" .....

Poznamenejme, že princip zúplnění z praxe pop-music přímo vnucuje založení Obchodní akademie múzických umění.

Snad v brzké době bude museum v moskevském Kremlu obohaceno o čapku Stereomachovu. Věřím, že třetí nádvoří Pražského hradu bude už konečně zdobit stereolit. V přehledu publikací se budou preferovat nejen monografie, ale také stereografie. ...

Z těchto ukázek vidíme, že princip zúplnění vede k plastičtějšímu pohledu na svět.

V rámci vědecké pravdy musím uvést, že idea principu zúplnění nepochází z mé hlavy. Mám v rukách pozůstalost svého dědečka *Jana N. Ullmanna* (• 24.5.1893 – + 2.11.1946), ve které je mnoho myšlenek, které ukazují cestu k principu zúplnění.

Již jako malý chlapec měl silně vyvinutý smysl pro spořivost. Ten se mj. projevoval v tom, že se snažil si šetřit nové hračky a až do omrzení si hrál se starými, i když byly do nemožnosti opotřebované. Jeho matka, která pocházela z Hané, jej často nabádala slovy: *"Jeníku, s tém si už nehré"* a ukazovala přitom na starou rozbitou hračku. Slovy: *"Nežinýruj se a hré si s tém"* mu podávala novou hračku. To činila soustavně, tak docílila, že Jeník se nakonec nežinýroval a hrál *"si s tém"* a hrál *"si s tém"* tak dlouho a neméně soustavně, až vytvořil *systém* a *systémové inženýrství* i položil základy teorie her dokonce dříve, než vznikly.

Teoretické poznatky právě v této oblasti her dovedl, jak uvádějí jeho zápisy i rodinná tradice, později neobyčejně úspěšně prakticky uplatnit.

Je známo, že za svého pobytu v Monte Carlu tímto způsobem získal závratné částky v různých hrách. Tyto prostředky mohl využít k financování svých nákladných projektů, aniž se musel uchylovat k pokoutnému zisku potřebných prostředků např. rozkrádáním majetku v kapitalistickém vlastnictví.

Musím uvést, že Jan N. Ullmann nezůstal u systémového inženýrství, ale vybudoval systémové doktorství (později rozvinuté jeho následníky až k systémovému kandidátství), systémové bohosloví, a zejména systémovou byrokracii[8].

Jan N. Ullmann v rámci boje za rovnoprávnost v odborné terminologii byl zastáncem rovnoprávnosti aplikace předpon:
mono-, bi-, di-, tri-, multi-, poly-, stereo-, kvadro-. Vrhl tím zcela nové světlo nejen na kvadraturu kruhu a trisekci úhlu, ale i např. na problematiku bigotnosti, bikavéru či dokonce trikotáže a stereogamie. K monoklu a binoklu hravě sestrojil stereokl a polykl. Některé otázky se mu nepodařilo vyřešit a jsou dodnes otevřené:

- "•" Kdy lze místo polykání monokat nebo alespoň bikat?
- "•" Je skutečně policajt polymerem monocajtu?
- "•" Je bidlo skutečně bidlo nebo pouze monodlo?
- "•" Jak potom vypadá polydlo?
- "•" Bizon je patrně pouze monozon.
- "•" Není trychtýř vlastně jen monochtýř?

Jan N. Ullmann tyto otázky nevyřešil, ale patří mu nehynoucí zásluha, že je nastolil.

Tyto kusé informace ukazují, že princip zúplnění dává nebývalé možnosti. Věřím, že se mi v této kusé informaci podařilo zachytit nejen jeho roli a aplikace, ale i část jeho prehistorie a historie.

[8]Vrátíme-li se zpět k principu zúplnění, tak v jeho světle zjistíme, že např. obecná teorie systémů se jeví pouze jako jedna z jeho dílčích a okrajových aplikací.

Ještě jednou o koeficientu determinace

Jan Ámos Víšek

O koeficientu determinace bylo jistě napsáno tak mnoho (viz [7],[8],[16] & všechny monografie věnované regresní analýze, viz např. [3] nebo [15]), že psát cosi dalšího není jen nošení dříví do lesa, ale prostě drzá troufalost. Tajně však doufám, že se najdou i tací, kteří velkoryse potlačí lítost nad dalším zmarněným papírem a dočtou tento příspěvek do konce.

Koeficient determinace je spolu se studentizovanými hodnotami odhadů regresních koeficientů a Fisher–Snedecorovou F-statistikou, jednou ze základních charakteristik klasické regresní analýzy. Je nasnadě dopátrat se, proč tomu tak je. O těchto charakteristikách se dá totiž pěkně přednášet – neboť jsou jednoduché, jejich rozdělení je radost odvozovat a koneckonců se to dá i dobře zkoušet studenty. Proto si tyto charakteristiky našly pevné místo v mnohých monografiích, počítačových knihovnách a přechodně v době zkoušek i v myslích některých studentů.

Tyto důvody pak patrně vedly k tomu, že pro příslušná rozdělení byly sestaveny tabulky a nalezeny (mimochodem opravdu těsné) aproximace (viz [1]) použitelné na počítačích (pokud ovšem došlo k tomu, že se nevysvětlitelným řízením osudu do implementované verze nevloudila chyba). Četnost zmínek o těchto rozděleních (t a F) může být snad překonána jen četností referencí na normální rozdělení. Toto, ač empiricky nerozlišitelné od t (a to i pro dosti malé počty stupňů volnosti), dovoluje odvození ještě pozoruhodnějších výsledků, které jsou všeobecně pokládány za rodinné stříbro statistiky, patrně také díky tomu, že se z učebnic decentně vytratilo varování Sira Ronalda Aylmera Fishera (viz [4]), že výkonnost těchto výsledků povážlivě klesá (asymptoticky i k nule) již při zmíněném t-rozdělení (o jiných rozděleních ani nemluvě).

Zmíněná jednoduchost uvedených statistik se však může ukázat stejně zrádná jako vyschlé dno sudoměřského rybníka. V Tabulce 1 jsou uvedena data, která byla nasimulována pomocí modelu

$$Y_i = 10 + 11\,X_1 + 12\,X_2 + \cdots + 20\,X_{10} + \varepsilon_i, \quad i = 1, 2, \ldots, 18, \tag{1}$$

kde vysvětlující proměnné jsou z intervalu $[-2, 2]$ a u každého desetirozměrného vektoru byla jedna souřadnice změněna o $+10$ či -10 (některé body „replikovány“ a „posunuty“ v různých souřadnicích; data lze tedy spíše považovat za cosi mezi simulovanými a umělými). Jako „náhodný

šum“ byla použita posloupnost $\{\varepsilon_i\}_{i=1}^{18}$ nezávislých n.v. s rozdělením $N(0, \sigma^2)$ se $\sigma = 0.2$ (tedy použita byla „jedna realizace“). „Zašumění“ je tedy spíše formální, což znamená, že „deterministické“ jádro modelu (1) platí téměř přesně, a tudíž by jej měla být hračka rozpoznat. Nakonec byla data kontaminována, a to tak, že byly body 2 a 12 poněkud změněny. Snadno se nahlédne, že kontaminace nastala díky asi 6 „překlepům“ při (hypotetickém) vkládání dat do počítače (což představovalo asi 1 000 úderů). Možná, že si v této chvíli položíte otázku, co u všech čertů vedlo k tomuto výběru dat. Následující poznámka je tedy určena těm, kteří jsou zvědaví na to, proč právě data tohoto charakteru mohou být pro hledání „protipříkladů“ v regresi zajímavá, ostatní nechť ji přeskočí (až ke značce ♡).

V roce 1979 R. A. Maronna, O. H. Bustos a V. J. Yohai [9] ukázali, že mezi řešeními soustavy rovnic

(2)
$$\sum_{i=1}^{n} \psi\left(Y_i - \sum_{j=1}^{p} X_{ij}\,\beta_j\right) X_{ik}\,w(X_i) = 0 \qquad\qquad k = 1, 2, \ldots, p$$

(kde $X_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$) existuje alespoň jedno, které má bod selhání nižší než $\frac{1}{p}$. Jinými slovy, pokud existuje pouze jedno řešení, potom (poněkud paradoxně) příslušný M-odhad má bod selhání, pro vyšší dimenze problému, nevyhnutelně poměrně nízký (a to si prosím povšimněte, že jde o vážený M-odhad, jemuž sudičky (viz např. Hampel, Krasker, Welsh atd., viz [5]) dali do vínku možnost potlačit ta data, která si ve faktorovém prostoru našla příliš individualistickou pozici). Známe-li tento výsledek, pak se můžeme ptát, co vlastně tuto skutečnost způsobuje. Připomeňme však nejprve, že bod selhání je dán jako nejmenší počet bodů m (dělený celkovým počtem bodů, řekněme n), které když změníme, zcela znehodnotíme odhad. Představme si „oblak“ regresních dat okolo počátku, která jsou „nekontaminovaná“. Nyní aby náš protihráč (příroda) znehodnotil M-odhad daný vztahem (1), udělá z m_1 bodů „leverage pointy“ v prvé souřadnici. Ty způsobí, díky faktoru X_{i1}, potíže v řešení první ze soustavy rovnic v (1), tj. v řešení

$$\sum_{i=1}^{n} \psi\left(Y_i - \sum_{j=1}^{p} X_{ij}\,\beta_j\right) X_{i1}\,w(X_i) = 0.$$

Přidělíme tedy těmto bodům malé váhy $w(X_i)$. Nyní „protihrající“ příroda změní m_2 bodů v druhé souřadnici, my jim zase dáme malé váhy (abychom se zbavili faktu, že tyto body – díky faktoru X_{i2}, vystupujícím

v (1) – znehodnotily odhad). Celkově tedy budeme nuceni dát malé váhy $\sum_{k=1}^{p} m_k$ bodům. Avšak $\sum_{k=1}^{p} m_k$ může být nejvýše $n-p-1$ (aby nám alespoň $p+1$ bodů určovalo „správný“ model, pokud se neznámo jak stane, že budou mít velké váhy). Odtud alespoň pro jedno k je $\frac{m_k}{n} \leq \frac{1}{p}$.

♡ Vraťme se však k našim datům, tady jsou:

TABULKA 1 *Data řídící se modelem (1)*

case	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
1	9.39	-0.74	0.15	-0.58	1.39	1.99	0.28	-0.75	1.12	0.08	165.2
2	-1.22	0.55	11.83	0.37	-0.13	0.08	31.58	0.21	1.29	-11.7	9.4
3	0.62	-10.2	1.44	0.04	0.93	-0.70	0.52	0.52	0.75	0.46	-42.2
4	-1.22	0.55	1.83	0.37	-0.13	0.08	11.58	0.21	1.29	1.66	289.4
5	-1.35	-1.37	-0.71	0.55	-1.42	0.22	0.39	10.12	2.00	0.69	200.4
6	-1.21	0.78	0.83	1.90	-1.43	-0.81	-0.08	-0.56	-11.4	1.95	-179.0
7	-1.35	-1.37	-0.71	10.55	-1.42	0.22	0.39	0.12	2.00	0.69	160.4
8	0.07	11.95	0.73	-0.55	-0.54	-0.87	-1.25	-1.46	-1.52	-0.45	49.0
9	0.03	-1.36	-1.39	-1.52	-11.1	-0.36	-0.45	1.05	1.10	1.49	-156.1
10	0.07	1.95	0.73	-0.55	-0.54	9.13	-1.25	-1.46	-1.52	-0.45	88.0
11	10.03	-1.36	-1.39	-1.52	-1.15	-0.36	-0.45	1.05	1.10	1.49	103.9
12	0.62	-0.18	1.44	0.04	0.93	-10.7	0.52	0.52	10.75	10.46	181.7
13	1.23	-1.04	0.65	-11.1	1.77	1.23	-1.08	-1.57	1.61	0.61	-93.7
14	1.23	-1.04	0.65	-1.10	1.77	1.23	-11.1	-1.57	1.61	0.61	-123.7
15	-0.73	-0.16	-1.57	0.92	-0.99	-1.90	0.45	-11.6	-1.18	-0.93	-294.4
16	1.10	-0.40	-8.29	-0.65	0.92	-1.40	-1.70	-0.30	-0.96	1.44	-131.5
17	-0.98	-1.14	1.23	-0.34	-0.46	1.30	-1.12	1.01	-1.42	11.68	216.3
18	0.61	1.59	-0.19	-0.52	0.48	-1.97	0.96	0.47	1.67	-11.2	-165.7

Diagonální prvky projekční (hat) matice jsou

TABULKA 2 *Diagonalní prvky projekční matice*

case	1	2	3	4	5	6	7	8	9
diag.	0.49	0.85	0.51	0.27	0.55	0.83	0.60	0.64	0.94

case	10	11	12	13	14	15	16	17	18
diag.	0.57	0.55	0.80	0.59	0.45	0.65	0.74	0.44	0.54

Vzhledem k tomu, že hodnota devátého prvku je .938 můžeme k tomuto bodu podujmout podezření. To však vzápětí zmizí, neboť zjistíme, že v učebnicích doporučovaná kritická (diagnostická) hodnota pro diagonální prvky je $\frac{2p}{n} = 1.22$, viz např [2], [3] či [15]. (V učebnicích bývá jen občas explicitně uvedeno, že diagonální prvek hat-matice překročí hodnotu 1 pouze tehdy, pokud je ve výpočtu chyba. Proto zdolá hodnotu 1.22 jen výjimečně. Možná, že někdo v tomto momentu namítne, že některé monografie doporučují nenechat bez povšimnutí všechna pozorování s diagonálním prvkem větším než 0.2, viz např. [6]. Tento požadavek v sobě implicitně zahrnuje snahu o vyváženost dimenze modelu a počtu dat. Vzhledem k tomu, že jsme

se z důvodu udržení rozumného rozsahu článku museli omezit na menší počet dat, musíme tento požadavek oslyšet. K problému se kratičce vrátíme na konci článku. Úplnější diskusi ke kritickým hodnotám pro diagonální prvky hat matice lze nalézt v [3].)

Suma sumarum máme málo dat a neradi bychom s nimi plýtvali, a tak nad hodnotou .938 protekčně přimhouříme oko (navíc jsouce v God-like-position, stejně víme, že bod 9 není kontaminovaný).

Po aplikaci nejmenších čtverců dostaneme tyto odhady koeficientů a příslušné P-hodnoty.

TABULKA 3 *LS-odhady koeficientů a příslušné P-hodnoty*

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
odhad	-5.82	10.67	11.56	1.916	14.39	14.11	20.28	9.240	18.71	16.29	21.26
P-hodnota	0.782	0.119	0.047	0.814	0.025	0.075	0.013	0.038	0.007	0.022	0.001

Koeficient determinace nabyl hodnoty .919, parametr měřítka byl odhadnut na $\hat{\sigma} = 75.3$ a součet čtverců residuí byl 39 674.8.

Pokusíme se vyloučit ty vysvětlující proměnné (regresory), které jsou indikovány jako nevýznamné na 5% hladině významnosti (tj. absolutní člen, X_1, X_3 a X_5) a přepočítáme model. Diagonální prvky hat matrice vypadají nyní takto:

TABULKA 4 *Diagonalni prvky projekční matice pro redukovaná data*

case	1	2	3	4	5	6	7	8	9
diag.	0.05	0.80	0.41	0.19	0.41	0.68	0.52	0.55	0.02

case	10	11	12	13	14	15	16	17	18
diag.	0.50	0.02	0.76	0.52	0.16	0.57	0.03	0.39	0.40

Hodnota $\frac{2p}{n}$ je nyní .776 a ta napovídá, že bod 2 by mohl být „leverage pointem“. Vyloučíme-
-li tento bod a znovu přepočítáme model, dostaneme

TABULKA 5 *Diagonalni prvky projekční matice po vypuštění bodu 2*

case	1	3	4	5	6	7	8	9
diag.	0.05	0.41	0.55	0.41	0.69	0.52	0.56	0.02

case	10	11	12	13	14	15	16	17	18
diag.	0.50	0.02	0.77	0.52	0.49	0.57	0.04	0.43	0.43

a

TABULKA 6 *LS-odhady koeficientů a příslušné P-hodnoty*

	β_2	β_4	β_6	β_7	β_8	β_9	β_{10}
odhad	10.43	10.34	22.42	16.84	16.52	20.29	18.15
P-hodnota	0.069	0.079	0.006	0.008	0.011	0.006	0.002

Z Tabulky 6 plyne, že patrně X_2 a X_4 jsou stále ještě nevýznamné. Po jejich vyloučení dostaneme odhady

TABULKA 7 *LS-odhady koeficientů a příslušné P-hodnoty*

	β_6	β_7	β_8	β_9	β_{10}
odhad	19.93	18.09	16.15	17.00	16.97
P-hodnota	0.026	0.012	0.027	0.029	0.007

tj. konečně jsme dosáhli modelu, ve kterém jsou všechny vysvětlující proměnné významné. Koeficient determinace poněkud poklesl na .736, ale i to je nad nepsanou, ale tradovanou magickou hranicí 60 %. Konečně pak $\hat{\sigma} = 98.45$.

A JE TO! (Ale blbě.)

(Nechoďte ještě spát, na rozdíl od televizního večerníčku tenhle pokračuje.)

Po aplikování odhadu – Least Trimmed Squares –

$$\hat{\beta}^{\mathrm{LTS}} = \underset{\beta \in R^{11}}{\operatorname{argmin}} \sum_{i=1}^{15} r^2_{(i:18)}(\beta)$$

kde $r^2_{(i:18)}(\beta)$ je i-tá pořádková statistika mezi $r_i^2(\beta) = \left[Y_i - \beta_1 - \sum_{i=1}^{10} X_{ij}\beta_{j+1}\right]^2$, $i = 1, 2, ..., 18$, dostaneme následující hodnoty odhadů

TABULKA 8 *LTS-odhady koeficientů a příslušné P-hodnoty*

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
odhad	10.13	11.01	12.03	12.98	14.02	15.01	15.93	17.01	18.03	18.97	19.99
P-hodnota	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Odhad škály je $\hat{\sigma} = .19$.

Závěrem by tedy bylo možné doporučit:

Používejte procedury s vysokým bodem selhání a odmítněte nejmenší čtverce! **Bohužel, toto doporučení je stejně zavádějící jako víra v samospasitelnost tržní ekonomiky.** Snadno se totiž ukáže, že dvě procedury s 50 % bodem selhání (Least Median of Squares a Least Trimmed Squares) dávají na některých datech ortogonální odhady, (viz [12]).

Shrňme tedy zdánlivě neradostnou bilanci:

„Selhání“ klasických právě tak jako (vysoce) robustních procedůr lze obecněji nahlédnout asi následovně (viz [13]). Statistikové, ale nejen oni, uplatňují při formulování východisek k budování té či oné teorie zpracování dat řadu naprosto přesvědčivých heuristických argumentů a požadavků (např. požadavek maximalizace věrohodnosti, minimalizace součtu čtverců, konsistence, maximalizace bodu selhání, minimalizace maximálního vychýlení (maximum je bráno přes některou množinu distribucí), požadavek eficience, minimalizace (gnostických) informačních

ztrát atd.) a vymýšlejí stále nové (viz např. [10], jinak by totiž nemohli dostat granty). Tyto heuristické požadavky, přeformulované do matematických kriterií, se samozřejmě (přímo) „přesublimují“ do dobrých vlastností výsledných procedůr (ostatně tyto vlastnosti jsou hodnoceny podle stejných kritérií, takže úspěch je zaručen, až na tu trapnou maličkost, že se cestou musíme dokázat vypořádat s různými technickými problemy). Avšak naděje, že toto vše nakonec zaručuje rozumnost numerického výsledku při aplikaci těchto procedůr na data, má stejně racionální opodstatnění, jako naděje na vstup České republiky do Evropské dvanáctky v příštím roce.

Zdá se tedy, že blíže k jakési pragmatické pravdě je závěr:

Na heuristiku, která stojí v pozadí té či oné metody, nelze spoléhat. Jednoduché charakteristiky kvality výsledku (např. vhodnosti odhadu regresního modelu) mohou být snadno zavádějící. Jejich fungování je totiž (často) podmíněno postačitelností a eficiencí příslušných statistik. Postačitelnost a eficience je však svázána s rozděleními, které neumíme (empiricky) odlišit od těch, pro která tyto statistiky jsou hrubě deficientní (viz [6] *či* [5]*).*

Vhodnost odhadu regresního modelu je tedy patrně nejlépe posuzovat všemi dostupnými kriterii současně, a zejména pak dle „globálního pohledu“ na residua, jak např. činí „normal plot“. Tato diagnostická pomůcka (ač patrně lepší než koeficient determinace či studentizované hodnoty odhadů koeficientů) má však nejméně dva nedostatky:

- Za prvé neprodukuje číselně, tj. „objektivně“ posouditelný test.
- Za druhé je vhodná jen pro normálně rozdělená rezidua.

Ten druhý nedostatek se dá částečně odstranit tím, že použijeme jiné než normální kvantily.

Lepší, ale nepoměrně složitější řešení je např. posoudit shodnost (či rozdílnost) rozdělení residuí v různých oblastech faktorového prostoru (prostoru vysvětlujících proměnných), viz [14]. (Pokud ovšem je už odladěn program na toto srovnání, je to rutina; budete-li jej chtít zastavte se pro něj u mne.)

Dvě malé poznámky na konec. Samozřejmě může vyvstat námitka, že výše uvedený příklad obsahoval příliš málo dat. Druhá námitka by mohla být taková, že kdybychom na originální data použili diagnostický postup (známý praktikům), totiž hledání bodu, jehož vyloučení způsobí největší změnu v odhadech koeficientů (viz např. [15]), postupně bychom

vyloučili bod 2 a poté bod 12. Tento postup (však) zafunguje právě díky tomu, že jsme měli malý počet dat.

Obě námitky lze spravit tím, že zreplikujeme data (a to tolikrát, aby odhadnutý model „seděl“ už jen na špatných datech, což nastane díky tomu, že body 2 a 12 jsou „mírné leverage pointy“) a uděláme v replikách menší modifikace. Takto získaná data se již nehodí pro prezentaci v článku (neboť je jich příliš mnoho a jsou nenázorná), ale jinak zafungují velmi podobně tomu, co bylo popsáno výše.

Na úplný závěr vážné slovo. Jistě cítíte, že některé formulace v článku byly poněkud, doufám však že nikoliv neúnosně nadsazeny. Doufám, že to u těch z nás, kteří se více zabýváme praxí, povede ke snaze zamyslet se nad klasickými nástroji statistiky a jejich (bohužel vrozenými) omezeními. U těch z nás, kteří více tíhneme k teoretickým hájemstvím, to snad vyvolá pocit, že k záplavě nových, zejména estimačních metod, by nebylo od věci tvořit také rozmanité, avšak relativně snadno aplikovatelné (a nejen asymptoticky účinné) diagnostické nástroje.

Reference

[1] Abramowitz, M., Stegun, I. A. (1964): *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Dover Publications.

[2] Antoch, J., Vorlíčková, D.: *Vybrané metody statistické analýzy dat.* Academia, Praha, 1992.

[3] Chatterjee, S., Hadi, A. S. (1988): *Sensitivity Analysis in Linear Regression.* New York: J. Wiley & Sons.

[4] Fisher, R. A. (1922): On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A 222, pp. 309–368.*

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986): *Robust Statistics – The Approach Based on Influence Functions.* New York: J.Wiley & Sons.

[6] Huber, P.J.(1981): *Robust Statistics.* New York: J.Wiley & Sons.

[7] Kozák, J. (1993): *Znovu ke koeficientu determinace*, Informační bulletin České statistické společnosti, srpen 1993, 12–16.

[8] Kozák, J. (1994): *STATGRAPHICS a koeficient determinace*, Informační bulletion České statistické společnosti, duben 1994, 16–20.

[9] Maronna, R.A., Bustos, O. H., Yohai, V. J. (1979): Bias- and efficiency-robustness of general Mestimators for regression with random carriers. In *Smoothing Techniques for Curve Estimation. Eds. T Gasser and M. Rosenblatt, New York: Springer-Verlag, 91 - 116.*

[10] Simpson, D. G., Ruppert, D., Carroll, R. J. (1992): On one-step GM estimates and stability of inferences in linear regression. *Journal of American Statistical Association, vol. 87, 439 - 450.*

[11] Rubio, A. M., Víšek, J. Á. (1994): Diagnostics of regression model: Test of goodness of fit, *Transactions of the Fifth Prague Symposium on Asymptotic Statistics, eds. M. Hušková & P. Mandl, Springer Verlag, 423 - 432.*

[12] Víšek, J. Á. (1994 a): High robustness and an illusion of truth. *Transactions of ROBUST'94, JČMF (Union of Czech Mathematicians), eds. J. Antoch & G. Dohnal, ISBN 80-7015-492-6, 172-185.*

[13] Víšek, J. Á. (1994 b): On the heuristics of statistical results, submitted to *Proceedings of 'PROBASTAT'94'.*

[14] Víšek, J.Á.(1994 c): Testing the fit of regression model, submitted to *Mathematical Methods of Statistics.*

[15] Zvára, K. (1989): *Regresní analýza.* Praha: Academia.

[16] Zvára, K. (1993): *Který model je ten pravý*, Informační bulletin České statistické společnosti, květen 1993, 8–11.

14

Znáhodněné dotazování

Martin Anděl

1. Úvod

Výzkumy, které se týkají choulostivých otázek způsobu života populace, hluboce zasahují do soukromí jednotlivých respondentů. Výsledky takovýchto výzkumů jsou pak ovlivněny vysokým počtem odmítnutých odpovědí a dále i jistým množstvím nepravdivých odpovědí. Respondenti, kteří neodpovídají podle skutečnosti, se za pravdivou odpověď stydí nebo se obávají následného trestního stíhání. Tyto překážky nejsou zdolány ani ujištěním respondenta o tom, že dotazování je anonymní a data budou využita pouze pro statistické vyhodnocení. V následující kapitole je uveden postup, který zaručí respondentovi anonymitu i před tazatelem. Postup se týká těch otázek, na které očekáváme odpověď typu ano - ne.

2. Warnerův model

Nechť je dán nějaký znak A, který dělí populaci lidí na dvě disjunktní skupiny. Skupinu lidí se znakem A označíme jako skupinu $\mathcal{A}$. Znakem A může být například to, zda si jedinec doma nelegálně pálí alkohol, nebo zda je jedinec homosexuálně orientovaný. Naším úkolem je odhadnout podíl p_A lidí patřících do skupiny $\mathcal{A}$.
Standardní postup vypadá následovně. Z populace náhodně vybereme n respondentů, kterým položíme otázku:

„Jste členem skupiny $\mathcal{A}$?“

Celkový počet kladných odpovědí označme n_A. Odhad pro podíl p_A je dán

$$\hat{p}_A = \frac{n_A}{n}. \tag{2.1}$$

Rozptyl tohoto odhadu je

$$var(\hat{p}_A) = \frac{p_A(1-p_A)}{n}. \tag{2.2}$$

Ve Warnerově modelu znáhodněného dotazování (viz [1]) jsou respondentovi předložena dvě tvrzení:

15

1. Jsem člen skupiny $\mathcal{A}$.
2. Nejsem člen skupiny $\mathcal{A}$.

Respondent si pomocí náhodného pokusu, jehož výsledek zůstane tazateli neznámý, vylosuje tvrzení, ke kterému se vyjádří. Není proto možné určit, potvrzuje-li respondent svým „ano" nebo „ne" příslušnost ke skupině $\mathcal{A}$, nebo ji naopak popírá. Zaveďme následující označení:

p_A = *skutečný podíl osob náležejících do skupiy A,*
P = *pravděpodobnost, že si respondent náhodným pokusem vybere 1. tvrzení,*
n = *rozsah výběru,*
m = *celkový počet kladných vyjádření na obě tvrzení,*
λ = *pravděpodobnost kladné odpovědi.*

Pravděpodobnost λ je dána vzorcem

$$\lambda = Pp_A + (1 - P)(1 - p_A). \tag{2.3}$$

Nahradíme-li v (2.3) λ odhadem $\hat{\lambda} = m/n$, dostaneme odhad pro podíl p_A

$$(\hat{p}_A)_W = \frac{1}{2P - 1}\left(P - 1 + \frac{m}{n}\right), \qquad P \neq \frac{1}{2}. \tag{2.4}$$

Rozptyl tohoto odhadu je

$$var((\hat{p}_A)_W) = \frac{p_A(1 - p_A)}{n} + \frac{P(1 - P)}{n(2P - 1)^2}. \tag{2.5}$$

První sčítanec v (2.4) je obyčejný binomický rozptyl spojený s přímou otázkou. Druhý sčítanec v (2.4) je cena, kterou zaplatíme za nejistotu spojenou se znáhodněnou odpovědí. Rozptyl v (2.4) klesá s rostoucí vzdáleností P od 0,5. Zvolíme-li však P hodně blízko 0 nebo 1, dojde opět k neochotě respondenta odpovídat podle pravdy. Zdá se, že vhodným kompromisem mezi minimalizací rozptylu odhadu a ochotou respondenta odpovědět podle pravdy je volba parametru P někde mezi 0,1 a 0,3 nebo mezi 0,7 a 0,9.

3. Pálíte si doma alkohol?

Warnerova metoda znáhodněného dotazování byla úspěšně aplikována Ing. Josefem Machkem CSc. a Huverem Fernándezem Rodríguezem v 80. letech na Kubě (viz [3]). Cílem šetření bylo odhadnout procento domácností, které si samy nelegálně vyrábějí alkohol. Šetření bylo

16

provedeno ve třech oblastech. Pro srovnání byla použita též metoda přímého dotazování. Výsledky jsou uvedeny v tabulce 3.1.

Tabulka 3.1

oblast	*přímé dotazování*	*náhodněné dotazování*	
	odhad	*rozsah výběru*	*odhad*
1	*9%*	*384*	*60%*
2	*13%*	*380*	*30%*
3	*23%*	*576*	*40%*

4. Dvě spolu nesouvisející otázky

Tato metoda je modifikací Warnerova postupu. Respondentovi je předložena dvojice tvrzení:

1. Jsem člen skupiny $\mathcal{A}$.
2. Jsem člen skupiny $\mathcal{Y}$.

Skupina $\mathcal{Y}$ je skupina lidí, která je charakterizovaná nějakým obyčejným, společensky nezávadným znakem. Zde budeme předpokládat, že je nám pravděpodobnostní zastoupení p_Y skupiny $\mathcal{Y}$ v populaci známo. Pro pravděpodobnost kladného vyjádření λ platí

$$\lambda = Pp_A + (1 - P)p_Y . \tag{4.1}$$

Nahradíme-li v (4.1) λ odhadem $\hat{\lambda} = m/n$, dostaneme odhad pro podíl p_A

$$(\hat{p}_A)_Y = \frac{1}{P}\left(\frac{m}{n} - p_Y(1 - P)\right) . \tag{4.2}$$

Rozptyl tohoto odhadu je

$$var((\hat{p}_A)_Y) = \frac{1}{nP^2}\lambda(1 - \lambda) . \tag{4.3}$$

Parametry P a p_Y je vhodné volit následovně. Parametr P by měl být blízko jedné, např. mezi 0,7 a 0,9. Volba parametru p_Y závisí na hodnotě odhadované pravděpodobnosti p_A. Očekáváme-li, že pravděpodobnost p_A bude menší než 0,5, snažíme se zvolit p_Y blízko 0. Očekáváme-li, že pravděpodobnost p_A bude větší než 0,5, snažíme se zvolit p_Y blízko 1. Je-li pravděpodobnost $p_A = 0,5$, snažíme se zvolit p_Y buď blízko 0 nebo blízko 1. Při volbě p_Y „blízko 0" nebo „blízko 1" je nutno vzít v úvahu, že rozptyl v (4.3) sice klesá s parametrem p_Y blížícím se k 0 nebo k 1, ale při

17

hodnotě velmi blízko 0 nebo 1 je ochota respondentů pravdivě odpovědět malá.

5. Odhad podílu homosexuálů v populaci ČR

V průzkumu firmy GfK Praha byla 889 respondentům starším 18 let položena následující dvojice otázek:

1. Jste homosexuálně orientovaný(á)?
2. Četl(a) jste včera (je-li pondělí, pak především) deník MLADÁ FRONTA DNES?

Tazatelé byli vybaveni šesti kartami s čísly 1, 2, 3, 4, 5 a 6. Po zamíchání nechali respondenta náhodně vytáhnout jednu kartu. Jestliže bylo na vytažené kartě číslo 2, pak měl respondent odpovědět na druhou otázku. V ostatních případech respondent odpovídal na první otázku. Četnosti odpovědí jsou uvedeny v tabulce 5.1.

Tabulka 5.1

odpověď	*četnost*
ano	*45*
ne	*730*
neodpověděl	*114*

Respondenty, kteří odmítli odpověď, z dalšího zpracování vynecháme. Hodnoty parametrů jsou následující:

$$P = \frac{5}{6}, \quad n = 775, \quad m = 45.$$

Z výzkumů čtenosti novin a časopisů převezmeme hodnotu

$$p_Y = 0,14$$

(viz [2]). Po dosazení do (4.2) dostaneme odhad pro podíl homosexuálů v populaci České republiky

$$(\hat{p}_A)_Y \doteq 0,042.$$

Dosadíme-li do (4.3) místo λ odhad $\hat{\lambda}$

$$\hat{\lambda} = P(\hat{p}_A)_Y + (1 - P)p_Y,$$

obdržíme odhad pro rozptyl

$$var((\hat{p}_A)_Y) \doteq 0,0001.$$

18

Zveřejňované výsledky výzkumů čtenosti novin a časopisů jsou váženy podle aktuálních demografických údajů Českého statistického ústavu. Hodnota $p_Y = 0,14$ *je nevážená. Vážená hodnota* p_Y *je 0,16. Odhad* $(\hat{p}_A)_Y$ *vypočtený s touto váženou hodnotou je 0,038.*
Výsledné hodnoty jsou v souladu s obecně uváděným 4% zastoupením homosexuálů v populaci. Společnost DEMA, která na zakázku pražského Sexuologického ústavu uskutečnila výzkum sexuálního chování občanů České republiky, však v Mladé frontě DNES dne 9. 3. 1994 uvedla, že podíl homosexuálně orientovaných mužů i žen je v naší populaci maximálně dvouprocentní.

LITERATURA

1. reenberg B. G., Abul-Ela A. A., Simmons W. R., Horvitz D. G. (1969): The unrelated question randomized response model: Theoretical Framework. *J. Amer. Statist. Soc.* **64**, 520–539.
2. edia projekt 94, AISA MEDIA, GfK PRAHA, SKMO. 3. kvartál 1. 8. - 30. 9. 1994 , Praha 1994.
3. odríguez H. F. (1984): Respuesta aleatorizada. Extension de la tecnica de Warner a modelos muestrales de uso frecuente en poblaciones humanas. Kandidátská disertační práce, Universidad de la Habana.

Autorova adresa: GfK Praha, spol. s r. o., Institut pro výzkum trhu, Újezd 40/450 (ČOS), 118 01 Praha 1.

SoftStat '95

Dan Pokorný

V lichých rocích, tedy právě v mrtvém mezidobí mezi dvěma COMPSTATy se v německém Heidelberku koná SoftStat, deklarovaný jako „konference o vědeckých aplikacích statistického software". Tématika konference je v mnohém příbuzná spektru Compstatu: Od matematicky orientaovaných přednášek v oblasti výpočtové statistiky ke komerční bitvě statistických softwarových velikánů. Spádovou oblastí SoftStatu byly původně německy mluvící země, nyní se však stává stále více událostí přinejmenším evropského významu s četnými zámořskými hosty. Odráží se to i v jazykové struktuře příspěvků přihlášených na lestošní konferenci, kdy angličtina převládá nad němčinou v poměru 87:15.
Tématické okruhy konference jsou:

19

- Výuka statistiky a statistický software
- Statistické poradenství
- Geografické informační systémy
- Škálování a klasifikace
- Statistické systémy (SPSS, S-PLUS, SAS, BMDP, SYSTAT, Statistica aj.)
- Statistické systémy a modelování
- Plánování pokusů
- Informační systémy v Internetu (nově zařazená, velmi aktuální oblast)
- Explorativní analýza dat
- Matadata a informační systémy
- Kvalitativní sociální výzkum
- Kvantitativní obsahová analýza
- Statistika a neuronové sítě
- Algoritmické aspekty statistické analýzy dat
- Simulace

SoftStat je organizován společností ZUMA (Zentrum fuer Umfragen, Methoden und Analysen) v Mannheimu, která v Německu šíří statitickou a metododickou osvětu na poli empirického výzkumu. Duši SoftStatu je pan Frank Faulbaum, který se v minulosti velmi zasloužil o neformální spolupráci mezi analytiky dat žijícími na území rozdělené Evropy.
Konference se koná ve dnech 26.-30. března v novém areálu University v Heidelberku,. Konferenční poplatek je DM 290.-, pro studenty je zlevněn na DM 50.-, od vystavovatelů komerčního software je naopak vyžadován příplatek DM 150.-. Před či po konferenci se lze za zvláštní poplatek zúčastnit jednoho ze dvou kurzů o systémech pro "structural equation modelling" (Joreskog: LISREL a Bentler: EQS).

Program a přihlášky:

ZUMA	Tel: 0049 621 1246-174
Postfach 12 21 55	Fax: 0049 621 1246-100
B 2,1	e-mail: sofstat@zuma-mannheim.de
D–68072 Mannheim	Gopher: gopher.social-science-gesis
Germany	

Turistické informace:

Heidelberk je podle mého soudu hezké město s zachovalým historickým jádrem, které stojí za to navštívit. Podobně jako Praha ma Karlův most, na kterém však skromně stojí jedinná socha a jednalo se o jiného Karla. Na vrch nad městem vedou dvě lanovky konstrukčně podobné té petřínské. Na zámku lze shlédnout největší vinný sud na světě, který byl naplněn a vypit jen jednou a stal se tak předchůdcem dnešních nevratných obalů. Březen lze pro návštěvu Heidelberku doporučit, neboť letní měsíce mají pravidelně zamluvené Američané a Japonci. Strategicky výhodné je ubytovat se v historickém středu města, které se ve večerních hodinách stává pro účastníky konference atraktorem.

pokorny@rzmain.rz.uni-ulm.de

20

Ze Společnosti

Stejně jak tomu bylo v minulých letech, i v tomto roce Vám zasíláme *složenku na zaplacení členského příspěvku* ve výši 60,– Kč, případně vyšším (dle možností nebo dle dluhů za minulá léta). Při placení prosím vyhovte našim prosbám, které jsme uváděli (a zdůvodňovali) již vloni:

(1) *uvádějte jako variabilní symbol rodné číslo* (několik členů má stejná příjmení – viz adresář),

(2) při platbách více členů jednou složenkou *zašlete seznam jmen*, nejlépe na adresu H. Řezankové (viz adresář).

Pro úplnost opakujeme číslo účtu u České spořitelny (kdyby někdo ztratil složenku): 8024551/0800.

A nyní k **hospodaření České statistické společnosti v roce 1994**. Příjmy z členských příspěvků byly i přes pokles členské základny (především z řad členů ČSÚ) vyšší než v předchozích letech (11410,-), neboť někteří členové ČStS platili své nedoplatky za minulá léta. Úroky z vkladových certifikátů činily 1836,– Kč a úroky z běžného účtu 43,– Kč. Spolu se zůstatkem z roku 1993 (21543,–) bylo v r. 1994 na straně příjmů 34832,– Kč.

K největším výdajům patřily opět platby spořitelně – včetně tisku složenek (494,–) a daně z úroků z vkladového certifikátu (483,–). Z dalších výdajů to byly náklady spojené se dvěma pohoštěními (na výroční konferenci na VŠE a na semináři v Olomouci) v částce 771,– Kč, smuteční kytice prof. Likešovi (502,–), nákup obálek (360,–), papíru (91,–), příjmových dokladů (11,–) a poštovné (843,-), což činí celkem 3555,–.

Do roku 1995 tak bylo převedeno 31 277,– Kč, z toho 25000,– na vkladových certifikátech, 6207,– na účtu u České spořitelny a 70,– Kč v hotovosti.

Hanka Řezanková

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání a jednou v roce v anglické verzi. Předseda společnosti: Prof. Ing. Václav Čermák, DrSc., VŠE Praha, nám. W. Churchilla 4, 130 67 Praha 3. ISSN 1210 – 8022
Redakce: Dr. Gejza Dohnal, Jeronýmova 7, 130 00 Praha 3. E-mail: dohnal at cspuni12.bitnet