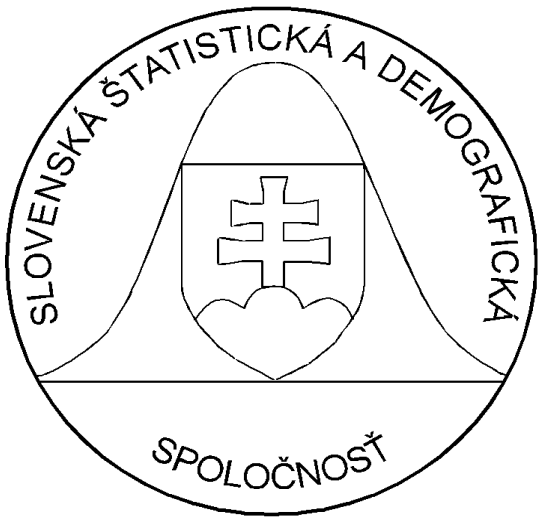
5/2010

FORUM STATISTICUM SLOVACUM



ISSN 1336-7420

Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava
www.ssds.sk

Naše najbližšie akcie:

(pozri tiež www.ssds.sk, blok Organizované akcie)

3. Medzinárodná konferencia Aplikácie metód na podporu rozhodovania „Inovácie“
9. 12. 2010 Bratislava, UM STU

MedStat, II. Celoslovenská konferencia medicínskej štatistiky
28. – 30. 1. 2011, Ružomberok

Pohľady na ekonomiku Slovenska 2011
12. 4. 2011, Bratislava, Aula EU

Nitrianske štatistické dni 2011
rok 2011, UKF Nitra

EKOMSTAT 2011, 25. škola štatistiky
29. 5. - 3. 6. 2011, Trenčianske Teplice

13. Slovenská demografická konferencia “Mesto a vidiek”
rok 2011, Nitriansky kraj

FERNSTAT 2011
rok 2011, UMB Banská Bystrica

20. Medzinárodný seminár Výpočtová štatistika
1. – 2. 12. 2011, Infostat Bratislava

Prehliadka prác mladých štatistikov a demografov
1. 12. 2011, Infostat Bratislava

Regionálne akcie
priebežne

Pokyny pre autorov

Jednotlivé čísla vedeckého recenzovaného časopisu FORUM STATISTICUM SLOVACUM sú prevažne tematicky zamerané zhodne s tematickým zameraním akcií SŠDS. Príspevky v elektronickej podobe prijíma zástupca redakčnej rady na elektronickej adrese uvedenej v pozvánke na konkrétne odborné podujatie Slovenskej štatistickej a demografickej spoločnosti. Názov word-súboru uvádzajte a posielajte v tvare: ***priezvisko_nazovakcie.doc***

Forma: Príspevky písané výlučne len v textovom editore MS WORD, verzia 6 a vyššia, písmo Times New Roman CE 12, riadkovanie jednoduché (1), formát strany A4, všetky okraje 2,5 cm, strany nečíslovať. Tabuľky a grafy v čierno-bielom prevedení zaradiť priamo do textu článku a označiť podľa šablóny. Bibliografické odkazy uvádzať v súlade s normou STN ISO 690 a v súlade s medzinárodnými štandardami. Citácie s poradovým číslom z bibliografického zoznamu uvádzať priamo v texte.

Rozsah: Maximálny rozsah príspevku je 6 strán.

Príspevky sú recenzované. Redakčná rada zabezpečí posúdenie príspevku oponentom.

Príspevky nie sú honorované, poplatok za uverejnenie akceptovaného príspevku je minimálne 20 €, poštovné 4 €. Za každú stranu naviac je poplatok 5 €.

Štruktúra príspevku: ***(Pri písaní príspevku využite elektronickú šablónu: http://www.ssds.sk/ v časti Vedecký časopis, Pokyny pre autorov.)***

Názov príspevku v slovenskom jazyku **(štýl Názov: Time New Roman 14, Bold, centrovať)**

Názov príspevku v anglickom jazyku **(štýl Názov: Time New Roman 14, Bold, centrovať)**

Vynechať riadok

Meno1 Priezvisko1, Meno2 Priezvisko2 (štýl normálny: Time New Roman 12, centrovať)

Vynechať riadok

Abstrakt: Text abstraktu v slovenskom (prípadne českom) jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

Abstract: Text abstraktu v anglickom jazyku, max. 10 riadkov (štýl normálny: Time New Roman 12).

Kľúčové slová: Kľúčové slová v slovenskom (prípadne českom) jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

Key words: Kľúčové slová v anglickom jazyku, max. 2 riadky (štýl normálny: Time New Roman 12).

JEL classification: Uviesť kódy klasifikácie podľa pokynov v:
<http://www.aeaweb.org/journal/jel_class_system.php>

Vynechať riadok a nastaviť si medzery odseku pre nadpisy takto: medzera pred 12 pt a po 3 pt. Nasleduje vlastný text príspevku v členení:

1. **Úvod** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať,)*
2. **Názov časti 1** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*
3. **Názov časti 1. . .**
4. **Záver** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

Vlastný text jednotlivých častí je písaný štýlom Normal: písmo Time New Roman 12, prvý riadok odseku je odsadený vždy na 1 cm, odsek je zarovnaný s pevným okrajom. Riadky medzi časťami a odsekmi nevynechávajte. Nastavte si medzi odsekmi medzeru pred 0 pt a po 3 pt.

5. **Literatúra** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, číslovať)*

[1] Písať podľa normy STN ISO 690

[2] GRANGER, C.W. – NEWBOLD, P. 1974. Spurious Regression in Econometrics. In: Journal of Econometrics, č. 2, 1974, s. 111 – 120.

Adresa autora (-ov): ***Uveďte svoju pracovnú adresu!!!*** *(štýl Nadpis 1: Time New Roman 12, bold, zarovnať vľavo, adresy vpísať do tabuľky bez orámovania s potrebným počtom stĺpcov a s 1 riadkom):*

Meno1 Priezvisko1, tituly1 (študenti ročník)
Pracovisko1 (študenti škola1)
Ulica1, 970 00 Mesto1
meno1.priezvisko1@mail.sk

Meno2 Priezvisko2 , tituly2 (študenti ročník)
Pracovisko2 (študenti škola2)
Ulica2, 970 00 Mesto2
meno2.priezvisko2@mail.sk

FORUM STATISTICUM SLOVACUM

vedecký recenzovaný časopis Slovenskej štatistickej a demografickej spoločnosti

Vydavateľ:
Slovenská štatistická a demografická spoločnosť
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Redakcia:
Miletičova 3
824 67 Bratislava 24
Slovenská republika

Fax:
02/39004009

e-mail:
chajdiak@statis.biz
jan.luha@fmed.uniba.sk

Registráciu vykonalo:
Ministerstvo kultúry Slovenskej republiky

Registračné číslo:
3416/2005

Evidenčné číslo:
EV 3287/09

Tematická skupina:
B1

Dátum registrácie:
22. 7. 2005

Objednávky:
Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava 24
Slovenská republika
IČO: 178764
DIČ: 2021504276
Číslo účtu: 0011469672/0900

ISSN 1336-7420

Ročník:
VI.

Číslo:
5/2010

Cena výtlačku: 20 EUR
Ročné predplatné: 80 EUR

ÚVOD

Vážené kolegyne, vážení kolegovia,

piate číslo šiesteho ročníka vedeckého recenzovaného časopisu Slovenskej štatistickej a demografickej spoločnosti (SŠDS) je zostavené z príspevkov, ktoré sú obsahovo orientované v súlade s tematikou 19. medzinárodného seminára Výpočtová štatistika a Prehliadkou prác mladých štatistikov a demografov. Tieto akcie sa uskutočnili v dňoch 2. a 3. decembra 2010 na Infostate v Bratislave. Organizátorom seminára bola SŠDS v spolupráci s Fakultou managementu UK v Bratislave, Infostatom, SAS Slovakia s.r.o. a klubom Dispersus.

Akcie, z poverenia Výboru SŠDS, zorganizoval Organizačný a programový výbor v zložení: Ing. Iveta Stankovičová, PhD. - predseda, doc. Ing. Jozef Chajdiak, CSc. – podpredseda, RNDr. Ján Luha, CSc. – tajomník, Mgr. Lukáš Pastorek – predseda Klubu Dispersus.

Novinkou je spolupráca s klubom Dispersus pri Prehliadke prác mladých štatistikov a demografov a spestrenie programu seminára o akciu klubu Dispersus pre mladých analytikov s názvom „Pohľady do analytiky - Analytika očami profesionálov“. Obsahom akcie bolo interdisciplinárne pásmo prednášok prezentátorov zo súkromnej, štátnej i akademickej sféry, ktorí vo svojej práci využívajú analytické metódy.

Recenziu príspevkov zabezpečili: Ing. Iveta Stankovičová, PhD., RNDr. Ján Luha, CSc., Ing. Ľubica Sipková, PhD. a Ing. Tomáš Želinský, PhD.

Toto číslo vedeckého časopisu FSS 5/2010 zostavili a pre tlač pripravili: Ing. Iveta Stankovičová, PhD. a RNDr. Ján Luha, CSc.

Veľmi nás teší neustály záujem o seminár Výpočtová štatistika. Výbor SŠDS oceňuje aktivitu mladých v rámci Prehliadky prác mladých štatistikov a demografov, čo svedčí tiež o dobrej práci pedagógov a ich študentov. Dúfame, že možnosť prezentácie zvyšuje odbornú úroveň mladých štatistikov a demografov.

Výbor SŠDS

Z HISTÓRIE SEMINÁROV VÝPOČTOVÁ ŠTATISTIKA
FROM THE HISTORY OF SEMINARS COMPUTATIONAL STATISTICS

Pri príležitosti 19. ročníka semináru Výpočtová štatistika uvádzame stručnú chronológiu predošlých ročníkov.

Prvý seminár sa uskutočnil 9. - 10. 12. 1986 z iniciatívy zamestnancov Katedry štatistiky VŠE v Bratislave a Katedry statistiky VŠE v Prahe zaoberajúcimi sa problematikou využitia výpočtovej techniky v riešení štatistických úloh. Príspevky účastníkov boli uverejnené v Informáciách SDŠS č. 3 a č. 4 v roku 1986.

Miestom konania Seminárov bola vždy budova Infostat-u a väčšina seminárov sa organizovala v spolupráci so Štatistickým úradom SR (resp. SŠU v Bratislave) a Infostat-om Bratislava (resp. VUSEIaR Bratislava).

Druhý seminár prebehol 8. 12. 1987, tretí seminár 11. - 12. 12.1990. Pád socializmu a spoločenské zmeny spôsobili určitú prestávku v organizácii seminárov Výpočtovej štatistiky.

4. seminár sa uskutočnil 7. - 8. 12. 1994.

Od 5. seminára uskutočneného 5. - 6. 12. 1996 sa už realizuje každoročne ako medzinárodný seminár.

6. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4.- 5. 12. 1997,
7. medzinárodný seminár Výpočtová štatistika sa uskutočnil 3. - 4. 12. 1998,
8. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 1999,
9. medzinárodný seminár Výpočtová štatistika sa uskutočnil 7. – 8. 12. 2000,
10. medzinárodný seminár Výpočtová štatistika sa uskutočnil 6. – 7. 12. 2001,
11. medzinárodný seminár Výpočtová štatistika sa uskutočnil 5. - 6. 12. 2002,
12. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4. - 5. 12. 2003,
13. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 2004,
14. medzinárodný seminár Výpočtová štatistika sa uskutočnil 1. - 2. 12. 2005,
15. medzinárodný seminár Výpočtová štatistika sa uskutočnil 7. - 8. 12. 2006,
16. medzinárodný seminár Výpočtová štatistika sa uskutočnil 6. - 7. 12. 2007,
17. medzinárodný seminár Výpočtová štatistika sa uskutočnil 4. - 5. 12. 2008,
18. medzinárodný seminár Výpočtová štatistika sa uskutočnil 3. - 4. 12. 2009,
19. medzinárodný seminár Výpočtová štatistika sa uskutočnil 2. - 3. 12. 2010.

Príspevky 2. seminára boli opublikované v Informáciách SDŠS č. 1/1989 a od 3. seminára sa publikujú v samostatnom Zborníku príspevkov príslušného seminára. Od 14. ročníka seminára sú už príspevky publikované vo vedeckom časopise vydávanom Slovenskou štatistickou a demografickou spoločnosťou s názvom FORUM STATISTICUM SLOVACUM.

Zameraním seminára je problematika na rozhraní počítačových vied a štatistiky. Tematické okruhy posledných seminárov sa nemenia:

- praktické využitie paketov štatistických programov,
- práca s rozsiahlymi súbormi údajov,
- vyučovanie výpočtovej štatistiky a príbuzných predmetov,
- praktické aplikácie výpočtovej štatistiky,
- iné.

V čase konania seminára Výpočtová štatistika sa uskutočňuje aj ***prehliadka prác mladých štatistikov a demografov***. Táto akcia prebieha od 7. seminára. Na 8. medzinárodnom seminári prezentovalo svoje práce 5 mladých štatistikov a demografov, na 9. medzinárodnom seminári už bolo 20 prác mladých štatistikov a demografov, na 10. bolo prihlásených 26 prác a na 11. bolo prihlásených 18 prác, ale vzhľadom na niekoľko prác vypracovaných skupinou autorov bol počet účastníkov vyšší než predošlý rok. Na 12. seminári bolo prihlásených 19 prác, pričom niektoré sú prácou viacerých autorov. Na ďalšom 13. seminári bolo prihlásených 9 prác od 12 autorov. V rámci 14. seminára bolo prihlásených 15 sólových prác mladých autorov. Na 15. seminári bolo prihlásených 20 prác mladých autorov. V rámci 16. seminára bolo prihlásených 17 sólových prác mladých autorov. V rámci 17. seminára bolo prihlásených 15 sólových prác mladých autorov. V 18. ročníku bolo prihlásených 12 sólových prác mladých autorov. V aktuálnom ročníku 19. seminára bolo prihlásených 15 prác mladých autorov.

Prípadní záujemcovia z radov mladých štatistikov a demografov (za mladých považujeme štatistikov a demografov pred ukončením vysokej školy, čiže nie už doktorandov) môžu získať informácie na *www.ssds.sk* , blok akcie a na e-mailových adresách:

chajdiak@statis.biz *jan.luha@fmed.uniba.sk* *iveta.stankovicova@fm.uniba.sk*

Informácie o najbližšom seminári získate na webovskej stránke SŠDS *http://www.ssds.sk/* resp. *http://www.statistics.sk* v bloku Slovenská štatistická a demografická spoločnosť.

Doc. Ing. Jozef Chajdiak, CSc.
STU Bratislava
predseda SŠDS

RNDr. Ján Luha, CSc.
LF UK Bratislava
vedecký tajomník SŠDS

Ing. Iveta Stankovičová, PhD.
FM UK Bratislava
predsedníčka Programového a organizačného výboru seminára Výpočtová štatistika

Srovnání váhových funkcí v M-odhadech a vliv konstant na výpočet odhadů v programu SAS 9.1.3

Comparison of Weight Functions in M-estimates and Impact of Constants on the Calculation of Estimates in SAS 9.1.3

Jindřich Černý, Dagmar Blatná

Abstract: The article focuses on M-estimates in robust regression models and its internal weight functions, especially in statistical packet SAS 9.1.3 while using ROBUSTREG procedure. It is shown curves of weight function used in SAS and its comparison with few comments to SAS. Also it is demonstrated how chosen constants in weight functions affect weight function and consequently M-estimates. There is also an example in the article.

Key words: M-estimate, outlier, leverage point, weight function, robust regression, SAS 9.1.3, ROBUSTREG procedure.

Klíčová slova: M-odhady, vlivné body, váhová funkce, robustní regrese, SAS 9.1.3, procedura ROBUSTREG.

JEL classification: C3,C2

1. Úvod

Dnes jsme svědky toho, že nejenom odborníci znalí statistiky přišli na výhody používání jiných regresních metod než jen metody nejmenších čtverců (MNČ). Když výzkumník narazí na reálný problém, zjistí, že MNČ není dostačující. Zejména předpoklady o normálním rozdělení chyb většinou nebývají dodrženy, vybočující pozorování ať již leverages (zásadně odlišné hodnoty ve vysvětlujících proměnných) či outliers (zásadně odlišné hodnoty ve vysvětlované proměnné) nebývají výjimkou, taktéž chybová rozdělení s „těžšími" konci nejsou u reálných dat nic neobvyklého. Nehodící se měření nebo body nelze jednoduše vypustit a stává se pak nutností použít některou metodu odolnou proti vybočujícím pozorováním, tedy metodu robustní.

Jednou z takových metod je M-regrese. Tato metoda používá pro odhady tzv. ρ funkce reziduí a z nich odvozené tzv. váhové funkce. Uvnitř těchto váhových funkcí se dále nacházejí konstanty, které značně ovlivňují průběh těchto funkcí a v konečném důsledku samotné odhady. Cílem tohoto článku je přiblížit některé váhové funkce a ukázat, jak zvolené konstanty ovlivňují jejich průběh, případně jak lze změnit konstanty, aby došlo k aproximaci jiných funkcí. Tyto problémy jsou řešeny v programu SAS 9.1.3, který obsahuje proceduru ROBUSTREG, která obsahuje M-odhady a má možnost volby váhových funkcí a změny jejich parametrů. Zároveň budou uvedeny konkrétní doporučení pro program SAS.

2. Popis M-odhadů (M-estimates) a váhových funkcí

M-odhady navrhl poprvé M. Huber v roce 1976. Při výpočtu M-odhadů viz [3] dochází k minimalizaci ρ funkce reziduí

$$\min \sum \rho(e) = \min \sum_{i=1}^{n} \rho(y_i - \sum_{j=1}^{p} x_{ij}\beta_{j-1}), \qquad (1)$$

kde p je počet vysvětlujících proměnných, k řešení této minimalizace lze použít derivace, pokud je ρ konvexní funkce, dostáváme jedno řešení pro odhady parametrů. Tato funkce

nemůže být zvolena libovolně (avšak může být nesymetrická!), musí splňovat alespoň následující požadavky

$$\rho(0) = 0\,, \tag{2}$$

$$\rho(e) \geq 0 \text{ pro všechna } e, \tag{3}$$

$$\rho(e_1) \geq \rho(e_2) \text{ pro všechna } |e_1| > |e_2|\,. \tag{4}$$

Protože odhady jsou citlivé na změnu měřítka, je používána jejich studentizovaná verze se směrodatnou odchylkou σ (scale) a studentizovaná verze derivace

$$\min \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_{j-1}}{\sigma}\right) \text{ resp. } \sum_{i=1}^{n} \psi\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\beta_{j-1}}{\sigma}\right) = 0\,, \tag{5}$$

protože tato směrodatná odchylka je většinou neznámá, musí se odhadnout také (v případě, že je známá, může být v proceduře zadána). Procedura ROBUSTREG program SAS používá pro odhad algoritmus nazývaný Iteratively Reweighted Least Squares (IRLS). Průběžně se odhaduje měřítko i M-odhady. Nejen pro účely tohoto algoritmu byla zavedena takzvaná váhová funkce w, která je definována jako

$$w(e) = \frac{\psi(e)}{e}, \tag{6}$$

program SAS verze 9.1.3 zná celkem 10 váhových funkcí, každá váhová funkce odpovídá funkci některé ρ funkci reziduí viz [5] a [1], tato volba přináší některé problémy. Přehled funkcí a jejich přednastavené konstanty jsou uvedeny v tabulce 1. Kromě Hamplovy a Mediánové funkce jsou přednastavené parametry spolu s přednastaveným parametrem odhadu měřítka (jedná se o iterační odhad pomocí mediánu) nastaveny tak, aby se M-odhady s 95% vydatností asymptoticky blížily normálnímu rozdělení.

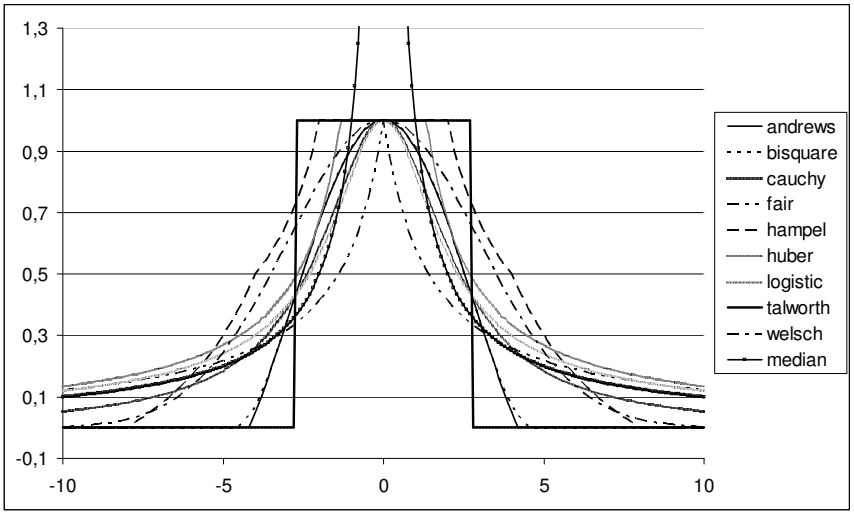
Tabulka 1: Váhové funkce a jejich nastavení

Název	$w(x)$	$\psi(x)$	$\rho(x)$	**pro x**	**Konst.**	
Andrews	$\frac{\sin(x/c)}{x/c}$	$c\sin\left(\frac{x}{c}\right)$	$c^2\left(1-\cos\left(\frac{x}{c}\right)\right)$	$\lvert x\rvert \leq \pi c$	c=1,339	(7)
	0	0	$2c^2$	*jinak*		
Bisquare	$\left(1-\left(\frac{x}{c}\right)^2\right)^2$	$x\left(1-\left(\frac{x}{c}\right)^2\right)^2$	$\frac{c^2}{6}\left(1-\left(1-\left(\frac{x}{c}\right)^2\right)^3\right)$	$\lvert x\rvert \leq c$	c=4,685	(8)
	0	0	$c^2/6$	*jinak*		
Cauchy	$\frac{1}{1+\left(\frac{\lvert x\rvert}{c}\right)^2}$	$\frac{x}{1+\left(\frac{x}{c}\right)^2}$	$\frac{c^2}{2}\ln\left(1+\left(\frac{x}{c}\right)^2\right)$	$x \in R^1$	c=2,385	(9)

Fair	$\frac{1}{1+\frac{\lvert x\rvert}{c}}$	$\frac{x}{1+\left(\frac{\lvert x\rvert}{c}\right)}$	$c^2\left(\frac{\lvert x\rvert}{c}-\ln\left(1+\frac{\lvert x\rvert}{c}\right)\right)$	$x\in R^1$	c=1,4	(10)
Hampel	1 $\frac{a}{\lvert x\rvert}$ $\frac{a}{\lvert x\rvert}\frac{c-\lvert x\rvert}{c-b}$ 0	x $a.\text{sign}(x)$ $\frac{a(x-c)}{b-c}$ 0	$\frac{x^2}{2}$ $c\lvert x\rvert-\frac{c^2}{2}$ $\frac{-acx}{b-c}+\frac{ax^2}{2(b-c)}$ $c^2/2$	$\lvert x\rvert\le a$ $a<\lvert x\rvert\le b$ $b<\lvert x\rvert\le c$ *jinak*	a=2 b=4 c=8	(11)
Huber	1 $\frac{c}{\lvert x\rvert}$	x $c.\text{sign}(x)$	$\frac{x^2}{2}$ $c\lvert x\rvert-\frac{c^2}{2}$	$\lvert x\rvert\le c$ *jinak*	c=1,345	(12)
Logistic	$\frac{\tanh(x/c)}{x/c}$	$c\tanh(x/c)$	$c^2\ln(\cosh(x/c))$	$x\in R^1$	c=1,205	(13)
Median	$\frac{1}{c}$ $\frac{1}{\lvert x\rvert}$	$\text{sign}(x)$	$\lvert x\rvert$	$x=0$ *jinak*	c=0,01	(14)
Talworth	1 0	x 0	$x^2/2$ $c^2/2$	$\lvert x\rvert\le c$ *jinak*	c=2,795	(15)
Welsch	$e^{\left(\frac{-1}{2}\left(\frac{x}{c}\right)^2\right)}$	$xe^{-\left(\frac{x}{c}\right)^2}$	$\left(\frac{c^2}{2}\right)\left(1-e^{-\left(\frac{x}{c}\right)^2}\right)$	$x\in R^1$	c=2,985	(16)

M-odhady nejsou robustní vůči outliers, avšak s vhodně zvolenou funkcí dodávají kvalitní výsledky pro leverages, zvláště máme-li k dispozici předpoklad o rozdělení chybové složky. Můžeme z něho metodou maximální věrohodnost dojít k $\rho(x)$ funkci reziduí. Například dojdeme-li k funkci $\rho(x)$ viz vzorec 7 tabulka 1, tak můžeme použít Andrewsovu váhovou funkci, viz tabulka 1. Pro přehlednost je zde uveden obrázek 1 všech použitých funkcí. Průběhy odpovídají deafultnímu nastavení konstant z programu SAS. Na tomto grafu je vidět, že některé funkce jsou velmi podobné a překrývají se, například Andrewsova a Bisquare se liší jen velmi nepatrně. Za povšimnutí stojí rovněž mediánová váhová funkce, z obrázku 1 je zřetelně vidět, že pro $x\to 0$ mediánová funkce překračuje hranici jedné a pro velmi malá x dosahuje vysokých hodnot. V [5] je uvedeno, že ji není doporučeno používat, protože odhady jsou nestabilní. To lze u některých datových souborů potvrdit, zvláště jsou-li

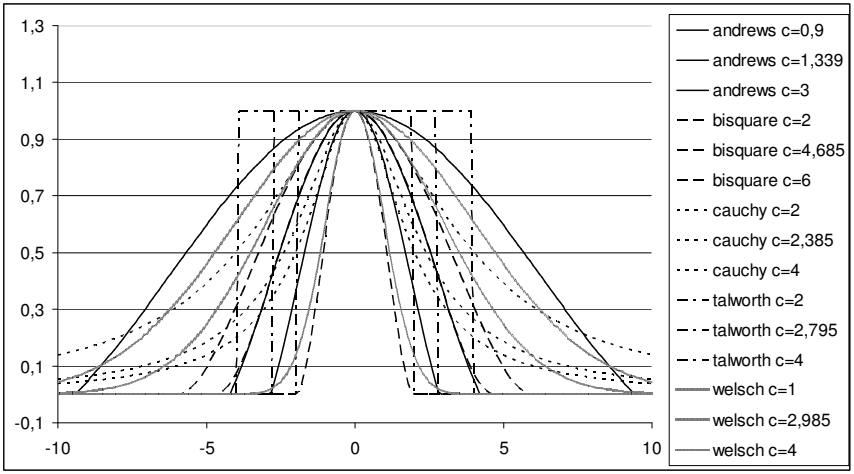
většího rozsahu, tato funkce je však velmi důležitá, neboť odpovídá, viz [1], Laplaceovu (též dvojité exponenciální) rozdělení chybové složky. Metodou maximální věrohodnosti tohoto rozdělení chybové složky dostaneme pro minimalizaci funkci absolutních reziduí, viz [2] K podobnému výsledku dojdeme také, kdybychom pro odhad parametrů použili tzv. L_p odhady. V tomto případě by se jednalo konkrétně o odhad v normě L_1. Problém při získávání těchto odhadů v programu SAS je způsoben použitou iterační metodou IRLS, tato metoda používá pro odhad, jak z názvu vyplývá, iterační postup s váhovou funkcí, tento iterační postup není příliš vhodný a při simulacích docházelo k selhání odhadů. Zde je na místě doporučení programátorům SASu zvolit pro tuto konkrétní funkci jinou vhodnější iterační metodu. Jinak program SAS provádí ostatní odhady bez větších problémů.



Obrázek 1: Srovnání přednastavených váhových funkcí

Vhodnou volbou konstant lze též průběhy funkcí přiblížit dalším váhovým funkcím, které nejsou v programu SAS obsaženy. Nevýhodou je, že přednastavení programu SAS neumožňuje volbu nesymetrických váhových funkcí. Na obrázku 2 je vidět změna průběhu váhových funkcí se změnou jejich koeficientů. Je zde pro přehlednost vybráno jen několik funkcí. Stejná váhová funkce pro různé koeficienty je vždy graficky znázorněna stejnou čarou. Vždy menší konstanta zašpičaťuje průběh váhové funkce. Pro váhové funkce použité v SASu platí, že se zmenšením konstant dostáváme robustnější odhady.

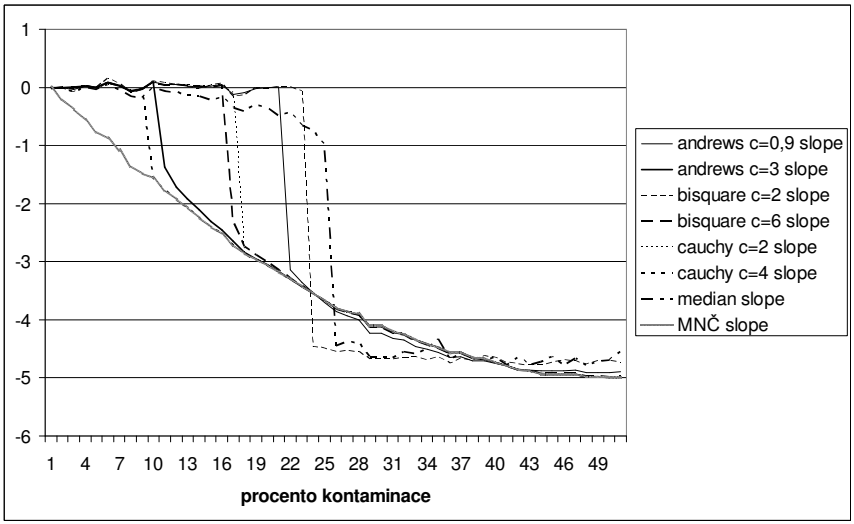
Zadaná konvergenční kritéria v proceduře dávají různě rychlé výsledky. Zvláště při vícerozměrných úlohách s velkým počtem pozorování může výpočet na běžném počítači trvat i několik minut (výjimečně i desítek minut). Nebyl však vypozorován vztah mezi výpočetní rychlostí (rychlostí konvergence) a zvolenou funkcí. Rychlost byla ovlivněna spíše strukturou dat. Další doporučení pro SAS 9.1.3 je nepoužívat lokalizovanou verzi, tato verze používá vnitřní názvy proměnných v češtině a poté ohlásí chybu, že těmto názvům proměnných nerozumí.



Obrázek 2:Změna průběhu váhových funkcí se změnou konstant

3. Příklady odhadů a selhání

Pro ukázku změny M-odhadů (ukázka je provedena pro ilustraci pouze s jednou vysvětlující proměnnou) byla použita kontaminace vysvětlované proměnné číslem, na této kontaminaci se nejlépe demonstruje okamžik selhání odhadů, viz obrázek 3, na kterém je vidět relativní změna směrnice při použití různých váhových funkcí s různými koeficienty. Pro odhad MNČ je vidět, že odhady selhávají při minimální kontaminaci. Je zde vidět i nestabilita odhadu s mediánovou váhovou funkcí, v jednorozměrné úloze však tato nestabilita není tak výrazná.



Obrázek 3:Selhání odhadů v závislosti na procentu kontaminace

Celkový počet provedených simulací byl v tomto případě 100, obrázek 3 ukazuje průměrné hodnoty. Kontaminace se při každé ze sta simulací vždy postupně zvětšovala. Hodnoty na ose x pak odpovídají kontaminaci v procentech. Byly prováděny simulace i s dalšími typy kontaminace a různou změnou nastavení dalších parametrů procedury, na jejich prezentaci zde však není prostor.

Výběr modelu se posuzuje robustním Akaikeho informačním kritériem AIC nebo Švarcovým kritériem a také pro M-odhady upraveným R^2. I testy jednotlivých parametrů a modelu jsou upraveny pro robustní verzi.

4. Závěr

V příspěvku byly popsány M-odhady v proceduře ROBUSTREG programu SAS 9.3.1 a váhové funkce používané v této proceduře. Bylo ukázáno, že podstatnější než volba funkce je nastavení vnitřních konstant těchto funkcí, toto bylo demonstrováno na provedených simulacích. Změnou těchto konstant lze nastavit míru robustnosti odhadů, což by mělo být vždy voleno s ohledem na použitá data.

M-odhady mají nezastupitelnou funkci všude tam, kde máme předpoklad nebo znalost rozdělení chybové složky reziduí, a dávají pak nejlepší možné odhady. Jak bylo však na začátku uvedeno, nejsou odolné proti outliers, i když toto neplatí vždy, avšak takovéto odhady jsou nestabilní, proto by se neměly v takovém případě používat.

M-odhady mají nezastupitelnou funkci všude tam, kde máme předpoklad nebo znalost rozdělení chybové složky reziduí, v takovém případě pak dávají nejlepší možné odhady. Jak ale bylo na začátku uvedeno, nejsou odolné proti outliers, i když to neplatí vždy, takové odhady jsou však potom nestabilní, a neměly by se proto v takovémto případě používat.

Poděkování: Tento příspěvek vznikl za podpory grantu IGA 26/2010

5. Literatura

[1] ANTOCH, J. – VORLÍČKOVÁ, D. 1992. Vybrané metody statistické analýzy dat. Praha: Academia 1992

[2] BRYNDÁK, M. 2001. Některé aspekty robustní regrese. Praha: VŠE 2001

[3] ČERNÝ, J. 2005. Srovnání vybraných metod robustní regrese. Praha: VŠE 2005

[4] HEBÁK, P. 1998. Regrese. Praha: VŠE, 1998.

[5] SAS 9.1.3 HELP, http://support.sas.com

Adresa autorů:

Jindřich Černý, Ing.
nám. W. Churchilla 4
130 67 Praha 3
Česká republika
cernyj@vse.cz

Dagmar Blatná, doc., Ing., CSc
nám. W. Churchilla 4
130 67 Praha 3
Česká republika
blatna@vse.cz

Measuring efficiency of the start-up and the survival of undertakings in selected countries of the European Union over time using data envelopment analysis

Meranie efektívnosti zakladania a prežitia podnikov vo vybraných krajinách Európskej únie pomocou dynamickej obalovej analýzy dát

Martin Boďa, Viera Roháčová

Abstract: The aim of the paper is to consider the easiness and successfulness of the starting-up and survival of undertakings in 21 member states of the European Union. The process of starting-up of undertakings is viewed as a result of the setting of the macroeconomic and microeconomic environment, and this holds for their survival, which allows to employ data envelopment analysis (DEA) for this purpose. A number of economic, legislative and bureaucratic variables playing a role in the setting-up of undertakings is considered as an inevitable input to the number of newly born enterprises as well as to the number of survived enterprises, which are taken as the outputs. The causal relationship between the inputs and the outputs is linked via data envelopment analysis and subjected to investigation so as to discover and measure efficiency of the process in question. As the conditions of the environment change over time, a specific approach is necessary to take account of these changes and motivates employment of dynamic data envelopment analysis. The results obtained permit to rank selected member states of the European Union in view of the incentives towards the entrepreneurial environment and to indicate areas in the entrepreneurial environment for a possible improvement.

Key words: undertakings, start-ups, survivals, efficiency, data envelopment analysis, window analysis, Malmquist index.

Abstrakt: Cieľom príspevku je posúdiť jednoduchosť a úspešnosť zakladania a prežitia podnikov v 21 členských štátoch Európskej únie. Proces zakladania podnikov a ich následného prežitia môže byť vnímaný ako výsledok určitých nastavení makroekonomického a mikroekonomického prostredia, čo umožňuje na tento účel použiť obalovú analýzu dát (DEA), ktorá je vo všeobecnosti chápaná ako metóda hodnotenia efektívnosti subjektov a je založená na hodnotení optimálnosti využitia vstupov a ich následnej transformácie do podoby výstupov. Vstupy v tomto prípade predstavujú vybrané ekonomické, legislatívne a byrokratické premenné, ktoré sú jedným z rozhodujúcich faktorov ovplyvňujúcich počet novozaložených a prežitých podnikov, ktoré možno v tomto prípade považovať za výstupy. Vzhľadom k skutočnosti, že stav celospoločenského prostredia sa v priebehu času mení, pokúsili sme sa aplikovať dynamickú obalovú analýzu dát, ktorá sa zameriava na meranie a vývoj efektívnosti v čase. Na základe získaných výsledkov je možné uskutočniť poradie vybraných členských krajín Európskej únie z hľadiska úspešnosti stimulov smerujúcich k podnikateľskému prostrediu a identifikovať oblasti možného zlepšenia efektívnosti.

Kľúčové slová: podniky, zakladanie podnikov, prežitie podnikov, efektívnosť, obalová analýza dát, window analýza, Malmquistov index.

JEL classification: C65, M13, C65

1. The introduction

Following the views of Joseph Schumpeter (see e. g. Pressman, 2006, pp. 160-2; Holman et al., 2005, pp. 273-4, 274-5), it is believed that the driving force of economic progress and growth is entrepreneurship, for which one of the outstanding traits is innovation and invention. The key factor that permits economies to grow and to thrive is a mass of entrepreneurs who are willing to accept risks and effect creative changes and improve on their technology or production, as if it were in the era of perfect competition. As Pressman (2006, p. 160) states of Joseph Schumpeters contribution, »when there were many entrepreneurs,

capitalism would thrive; on the other hand, if the entrepreneurial spirit was destroyed or severely hindered, capitalism would quietly transform itself into socialism«. The globalization of entrepreneurship, the amassment and internationalization of capital moves the internal structure of markets towards more rigid forms of imperfect competitions. Without any supporting arguments, it is doubtless that prevalence of transnational corporations and internationalized large enterprises would further suppress creative innovative and inventive spirit in the economic environment and the number of small and medium-sized enterprises would further diminished. The impacts would be upon welfare and prosperity.

Small and medium-sized undertakings, as the true and preservable agent of innovation, may be thus seen as a necessary condition for economic growth and prosperity. It is therefore in the best interest of every state to incite the will to undertake through variables in the macroeconomic and microeconomic environment. The entrepreneurship can thus be deemed as the black box of a transformation process in which the inputs are state-regulated settings in the economic environment that prompt or dissuade to engage in entrepreneurial activities and in which the outputs are the number of newly born (and surviving) undertakings. In this framework, it is of note to evaluate efficiency of the process of the start-up and survival of undertakings according to the pre-set factors in the economic environment, and to judge its change over time

Given the data available, the task of the evaluation of efficiency of the start-up and survival of undertakings, and its change over time, was accomplished in view of 21 selected countries of the European Union through the years 2004 – 2007. For the purpose of efficiency and efficiency change measurement, window analysis (under data envelopment analysis concept) was employed and changes in efficiency were computed and disaggregated by the Malmquist total efficiency index analysis.

2. The material

The data for the period of 2004 – 2007 entering the analysis were obtained from the web-site of the Eurostat and from the Doing Business Reports of Palgrave Macmillan and of the World Bank. The data set comprised the information on the following 21 countries: Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, Latvia, Lithuania, the Netherlands, Norway, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom. The analysis ran in two separate evaluations: for the model of start-ups and for the model of survivals. The choice of input and output variables for the two separate, yet related, analyses is shown in Table 1. It must be brought into attention that

- it was not meaningful to combine the two analyzed models considered into one for each model enables to assess a different aspect of support of small and medium-sized undertakings, the former evaluates the incentives to arise, and the latter the incentives to survive;
- the variables were selected so as not to reflect the size of economy (or state), in particular, the outputs (the number of undertakings newly born and the number of undertakings surviving) were expressed as a ratio to the number of active undertakings;
- the output variables are chosen to be maximized, and the input variable are to be minimized.

3. The methodology and the results

With respect to the selected outputs and inputs, it appears reasonable to assume variable returns to scale with the input orientated data envelopment analysis model: it is obviously possible to control the input variables (and minimize them) in order to achieve a higher rate of start-ups and survivals. This, of course, is under the competence of the state. A radial DEA

model, the BCC-I model, was chosen for window analysis, which was performed in the DEA-Solver learning version 3.0 application. In window analysis windows of length 1 trough 4 were analysed, which made possible to assess stability of findings between respective window analyses, i. e. to judge sustainability of efficiency dynamically over time. The results summarized for window analysis of window length 1 and window length 4 are displayed in Table 2, which also shows the Pearson and Spearman correlation coefficients between efficiency scores in respective years. For instance, the first Pearson correlation of 0.943 in the model of start-ups of undertakings measures the correlation between 2004 efficiency scores of length 1 window analysis and length 4 window analysis, and the similar holds for the Spearman rank correlation coefficient of 0.960. The remaining results can be supplied with the authors on request. It is notable that the Pearson correlation coefficients between average efficiency scores in the two models are 0.363 for length 1 window analysis and 0.231 for length 4 window analysis.

Table 1. The variables considered and their origin (The source: the authors.)

Variable	Measurement unit	Source
Output for the model of start-ups		
The number of undertakings newly born this year relative to the number of active undertakings this year	Proportion	Eurostat*†‡
Inputs for the model of start-ups		
The number of procedures necessary for the start-up	Number	Doing Business Reports of Palgrave Macmillan & World Bank*
The time necessary for the start-up	Number of calendar days	
The start-up costs relative to income per capita	Proportion	
The start-up minimum capital relative to income per capita	Proportion	
The rigidity of employment	Index	
The costs of redundancy of employees	Number of weeks of salary	
Output for the model of survivals		
The number of undertakings newly born last year surviving this year relative to the number of active undertakings this year	Proportion	Eurostat*†‡
Inputs for the model of survivals		
The rigidity of employment	Index	Doing Business Reports of Palgrave Macmillan & World Bank*
The costs of redundancy of employees	Number of weeks of salary	
The number of procedures necessary to enforce contracts	Number	
The time necessary to enforce contracts	Number of calendar days	
The costs necessary to enforce contracts relative to claim	Proportion	

* For all the years considered in the study, viz. 2004 – 2007, and for Austria, Bulgaria, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, Latvia, Lithuania, the Netherlands, Norway, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and the United Kingdom.

† Originally not available for Austria, Denmark, Germany, Norway and Portugal for 2004. The number of last-year born undertakings survived in 2004 was computed as an average of the linear prediction based on the values for 2005 – 2007 and of their mean.

‡ Calculated by the authors.

The Malmquist total efficiency index was computed and disaggregated according to Coelli (1996), using the DEAP program version 2.1, and in this fashion it gives notion of [1.] technical efficiency change (relative to the constant returns to scale), [2.] technological change, [3.] pure technical efficiency change (relative to the variable returns to scale), [4.] scale efficiency change, and [5.] total efficiency change. It holds that [1.] = [3.] × [4.] and [5.] = [1.] × [2.], or that [5.] = [2.] × [3.] × [4.], viz. total efficiency change can be decomposed into technological change, pure technical efficiency change and scale efficiency change. The disaggregation was calculated for all subsequent two-year periods (2004-2005,

2005-2006, 2006-2007), and for the entire four-year span (2004-2007). Table 3 reports only the results for both models for the entire four-year period of 2004-2007. The remaining results can be obtained from the authors on personal request.

Table 2. The results of window analysis (The source: the authors.)

The model of start-ups of undertakings

	Efficiency scores / Length of window = 1						Efficiency scores / Length of window = 4					
EU member state	**2004**	**2005**	**2006**	**2007**	**Average**	**Rank**	**2004**	**2005**	**2006**	**2007**	**Average**	**Rank**
Austria	0.642	0.624	0.586	0.578	0.608	15	0.547	0.555	0.568	0.578	0.562	15
Bulgaria	0.466	0.469	0.463	0.680	0.520	18	0.404	0.419	0.448	0.634	0.476	20
Czech Republic	0.769	0.769	0.525	0.526	0.647	14	0.700	0.700	0.516	0.526	0.611	12
Denmark	1.000	1.000	1.000	1.000	1.000	1	0.896	0.932	1.000	1.000	0.957	6
Estonia	0.914	0.625	0.824	0.679	0.761	11	0.757	0.573	0.737	0.658	0.681	11
Finland	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
France	0.619	1.000	1.000	1.000	0.905	9	0.583	1.000	1.000	1.000	0.896	8
Germany	0.479	0.477	0.475	0.478	0.477	21	0.456	0.456	0.471	0.478	0.465	21
Hungary	0.721	0.721	0.608	0.608	0.665	13	0.608	0.608	0.608	0.608	0.608	13
Italy	1.000	1.000	1.000	1.000	1.000	1	0.981	0.988	0.994	1.000	0.991	5
Latvia	0.856	0.805	0.775	0.791	0.807	10	0.727	0.739	0.771	0.791	0.757	10
Lithuania	1.000	1.000	1.000	1.000	1.000	1	0.709	1.000	1.000	1.000	0.927	7
Netherlands	0.702	0.690	0.590	0.681	0.666	12	0.573	0.574	0.576	0.671	0.599	14
Norway	0.943	0.888	0.897	1.000	0.932	8	0.809	0.814	0.883	1.000	0.877	9
Portugal	0.473	0.451	0.466	0.940	0.583	16	0.427	0.423	0.438	0.896	0.546	17
Romania	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	0.998	1.000	1.000	4
Slovakia	0.556	0.584	0.596	0.591	0.582	17	0.505	0.531	0.578	0.591	0.551	16
Slovenia	0.548	0.533	0.536	0.453	0.518	19	0.523	0.526	0.534	0.453	0.509	18
Spain	0.508	0.494	0.496	0.498	0.499	20	0.487	0.490	0.494	0.498	0.492	19
Sweden	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
United Kingdom	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Pearson correlations between efficiency scores for length 1 and 4							0.943	0.987	0.997	0.998	0.992	
Spearman correlations between efficiency scores for length 1 and 4							0.960	0.982	0.981	0.999	0.971	

The model of survivals of undertakings

	Efficiency scores / Length of window = 1						Efficiency scores / Length of window = 4					
EU member state	**2004**	**2005**	**2006**	**2007**	**Average**	**Rank**	**2004**	**2005**	**2006**	**2007**	**Average**	**Rank**
Austria	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Bulgaria	0.755	0.755	0.747	0.792	0.762	20	0.737	0.737	0.737	0.779	0.748	20
Czech Republic	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	0.972	1.000	0.993	9
Denmark	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Estonia	0.911	0.960	0.838	1.000	0.927	16	0.825	0.878	0.825	0.878	0.852	19
Finland	1.000	1.000	1.000	1.000	1.000	1	0.970	0.982	1.000	1.000	0.988	10
France	0.946	0.949	0.899	0.910	0.926	17	0.895	0.895	0.895	0.895	0.895	16
Germany	0.956	0.954	0.925	0.918	0.938	14	0.940	0.931	0.916	0.910	0.924	14
Hungary	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Italy	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Latvia	0.971	0.996	1.000	1.000	0.992	13	0.929	0.949	1.000	1.000	0.970	13
Lithuania	0.979	1.000	1.000	1.000	0.995	12	0.962	0.976	1.000	0.995	0.983	11
Netherlands	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Norway	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Portugal	0.929	0.926	0.895	1.000	0.938	15	0.907	0.898	0.871	0.913	0.897	15
Romania	1.000	1.000	1.000	1.000	1.000	1	0.985	1.000	0.994	0.922	0.975	12
Slovakia	0.885	0.886	0.932	0.923	0.907	18	0.857	0.859	0.919	0.897	0.883	17
Slovenia	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Spain	0.721	0.733	0.752	0.751	0.739	21	0.716	0.718	0.737	0.728	0.725	21
Sweden	0.885	0.885	0.893	0.867	0.883	19	0.860	0.860	0.868	0.867	0.864	18
United Kingdom	1.000	1.000	1.000	1.000	1.000	1	1.000	1.000	1.000	1.000	1.000	1
Pearson correlations between efficiency scores for length 1 and 4							0.970	0.968	0.996	0.904	0.978	
Spearman correlations between efficiency scores for length 1 and 4							0.958	0.943	0.942	0.812	0.935	

Table 3. The results of the Malmquist index disaggregation (The source: the authors.)

The disaggregation of total efficiency change measured by the Malmquist index												
	The model of start-ups						The model of survivals					
	Efficiency score changes - 2004 - 2007						Efficiency score changes - 2004 - 2007					
EU member state	technical efficiency change	techno-logical change	pure technical efficiency change	scale efficiency change	total efficiency change	Rank	technical efficiency change	techno-logical change	pure technical efficiency change	scale efficiency change	total efficiency change	Rank
Austria	0.903	1.138	0.901	1.002	1.027	16	0.975	0.906	1.000	0.975	0.884	17
Bulgaria	1.471	1.215	1.457	1.009	1.787	3	1.090	0.951	1.048	1.040	1.036	13
Czech Republic	0.651	1.100	0.684	0.951	0.715	21	0.974	0.994	1.000	0.974	0.968	15
Denmark	1.000	1.374	1.000	1.000	1.374	8	1.000	1.266	1.000	1.000	1.266	5
Estonia	0.711	1.190	0.743	0.957	0.846	20	1.103	0.999	1.098	1.004	1.102	7
Finland	0.958	1.282	1.000	0.958	1.228	10	1.317	0.995	1.000	1.317	1.311	4
France	2.075	1.227	1.615	1.285	2.546	1	1.019	1.072	0.961	1.060	1.092	9
Germany	0.798	1.231	0.998	0.799	0.982	18	0.978	0.896	0.960	1.019	0.877	19
Hungary	0.717	1.247	0.843	0.851	0.894	19	0.914	0.905	1.000	0.914	0.827	20
Italy	1.147	0.981	1.000	1.147	1.126	14	1.000	1.064	1.000	1.000	1.064	10
Latvia	0.857	1.255	0.924	0.927	1.075	15	0.950	1.099	1.030	0.923	1.044	12
Lithuania	1.060	1.340	1.000	1.060	1.421	6	2.269	1.623	1.022	2.221	3.681	1
Netherlands	1.319	1.323	0.970	1.359	1.745	4	1.147	0.961	1.000	1.147	1.102	7
Norway	1.095	1.266	1.060	1.032	1.386	7	1.000	0.879	1.000	1.000	0.879	18
Portugal	1.998	1.189	1.986	1.006	2.376	2	1.083	0.899	1.077	1.006	0.974	14
Romania	1.000	1.252	1.000	1.000	1.252	9	1.000	0.815	1.000	1.000	0.815	21
Slovakia	1.334	1.092	1.062	1.257	1.457	5	1.499	0.913	1.043	1.437	1.369	3
Slovenia	1.055	1.128	0.827	1.275	1.190	12	1.000	1.492	1.000	1.000	1.492	2
Spain	0.961	1.031	0.979	0.982	0.991	17	1.177	0.896	1.041	1.131	1.054	11
Sweden	0.866	1.312	1.000	0.866	1.136	13	1.080	1.040	0.980	1.102	1.124	6
United Kingdom	1.000	1.194	1.000	1.000	1.194	11	1.000	0.937	1.000	1.000	0.937	16
Mean	1.044	1.204	1.019	1.025	1.256		1.169	1.049	1.005	1.164	1.227	
Pearson correlations between efficiency changes of both models							0.081	0.232	0.176	0.194	0.081	
Spearman correlations between efficiency changes of both models							0.644	0.212	-0.150	0.631	0.290	

4. The discussion

The results in Table 2 are suggestive of the fact that efficiency development can be considered stable. There are not substantial differences in the efficiency scores for both models and for the countries considered when reviewing the results on length 1 window analysis and those on length 4 window analysis (and this is true for the other lengths). The impressions gained from the visual inspection of results are confirmed by the values of the correlation coefficients. However, some differences are seen (and confirmed by the correlation coefficients) between efficiency of the start-up of survivals and the efficiency of the survivals of undertakings in the countries studied.

Whilst from the point of view of the starting-up of undertakings, the most efficient countries comprise Denmark, Finland, Italy, Lithuania, Romania, Sweden and the United Kingdom. In the group of the most inefficient countries are found Bulgaria, Germany, Slovenia and Spain.

When assessing efficiency of the survival of undertakings, amongst the most efficient countries Austria, the Czech Republic, Denmark, Finland, Hungary and Italy rank. The most inefficient countries are assessed the Netherlands, Norway, Slovenia and the United Kingdom.

Window analysis facilitated the evaluation of the level and stability of efficiency over the entire period. Conversely, the Malmquist index analysis evaluates its progress over time. The results in Table 3 are suggestive that, in terms of the start-up of undertakings, the countries that worsened its position of 2004 through 2007 include the Czech Republic,

Estonia, Germany, Hungary and Spain (the Malmquist index lower than 1). From the standpoint of the survival of undertakings, the diminishment of efficiency was found with Austria, the Czech Republic, Germany, Hungary, Norway, Pourtugal, Romania and the United Kingdom. Though the Czech Republic, Germany and Hungary all indicated a general decrease of efficiency, the correlation coefficients reveal that these results do not come compatible at all.

The visual examination of individual efficiency scores gives the notion of the sources of the development of efficiency.

5. The conclusion

The article demonstrated the possibility of employment of window analysis and the Malmquist index analysis in the assessment of efficiency of the start-up and of the survival of undertakings. This possibility is grounded in the idea that the entrepreneurship is an innovative transformation process of inputs into outputs. It is purported that the state can control certain variables in the environment by which it can incite or obstruct entrepreneurship, regardless of whether it is the starting-up of undertakings or their surviving.

The article was prepared under the grant of VEGA No. 1/0828/10 Dynamic modelling of economic processes.

The bibliography

[1] COELLI, T. J. 1996. A guide to DEAP version 2.1: a data envelopment analysis (computer) program. In: Centre for Efficiency and Productivity Analysis (CEPA) Working Papers No. 8/96. Armidale (Australia): Universtiy of New England, Department of Econometrics, 1996. 50 pp. ISSN 1327-435X. ISBN 1-86389-4969.

[2] COOPER, W. W. et al. 2007. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-Solver software. 2nd ed. New York: Springer, 2007. 490 pp. ISBN 0-387-45281-8.

[3] HOLMAN, R. et al. 2005. Dějiny ekonomického myšlení. 3rd ed. Praha: C. H. Beck, 2005, 539 pp. ISBN 80-7179-380-9.

[4] PRESSMAN, S. 2006. Fifty major economists. 2nd ed. Oxon (GB): Routledge, 2006. 322 pp. ISBN 0-415-36649-6.

The authors' address

Martin Boďa, Ing. et Bc.
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta / Fakulta prírodných vied
KKMIS / KM
Tajovského 10 / Tajovského 40
975 90 / 974 01 Banská Bystrica
martin.boda@umb.sk

Viera Roháčová, Ing.
Univerzita Mateja Bela v Banskej Bystrici
Ekonomická fakulta
KKMIS
Tajovského 10
975 90 Banská Bystrica
viera.rohacova@umb.sk

Small-sample inference in the general Gaussian linear regression model under the assumption of multiplicative heteroskedasticity based on maximum likelihood estimation

Inferencia v zovšeobecnenom gaussovskom lineárnom regresnom modeli v malých výberoch za predpokladu multiplikatívnej heteroskedasticity založená na metóde maximálnej vierohodnosti

Martin Boďa

Abstract: In the article, the applicability and performance of three methods for the construction of a confidence region of a linear combination of the regression parameters in the general Gaussian linear regression model with the assumed multiplicative form of heteroskedasticity based upon the maximum likelihood estimation is investigated. A simulation study on the performance of these methods in small samples under several designs is submitted, when multiplicative heteroskedasticity is assumed whilst the true form may be additive, mixed, multiplicative, or "indefinite" heteroskedasticity.

Key words: Kenward-Roger method, additive heteroskedasticity, mixed heteroskedasticity, multiplicative heteroskedasticity, ML estimation, simulations.

Abstract: V článku sa vyšetruje využiteľnosť troch metód konštrukcie konfidenčnej oblasti lineárnej kombinácie regresných parametrov v zovšeobecnenom gaussovskom lineárnom regresnom modeli s predpokladanou multiplikatívnou formou heteroskedasticity založených na metóde maximálnej vierohodnosti. Predkladá sa simulačná štúdia o výkonnosti týchto troch metód v malých výberoch pri rôznych scenároch empirického skúmania, pričom sa predpokladá multplikatívna heteroskedasticita, hoci v skutočnosti môže ísť o aditívnu, zmiešanú, multplikatívnu alebo "neurčitú" heteroskedasticitu.

Key words: Kenwardova-Rogererova metóda, aditívna heteroskedasticita, zmiešaná heteroskedasticita, multiplikatívna heteroskedasticita, metóda maximálnej vierohodnosti, simulácie.

JEL classification: C13, C20.

1. The introduction

The considerations of the article are given to the general Gaussian linear regression model $(\mathbf{Y}, \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, in which $\mathbf{Y}$ is a Gaussian $(n \times 1)$ random vector with the expectation $\mathbf{X}\boldsymbol{\beta}$ and the covariance matrix $\boldsymbol{\Sigma}$. The design matrix $\mathbf{X} = (\mathbf{x}_1 \mid ... \mid \mathbf{x}_n)'$ is a non-random known $(n \times k)$ matrix of full column rank, and $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of unknown regression parameters. The $(n \times n)$ covariance matrix $\boldsymbol{\Sigma}$ is diagonal with multiplicatively heteroskedastic variance components in the form $\boldsymbol{\Sigma}(\boldsymbol{\alpha}) = \text{diag}\{ \exp(\mathbf{z}_1'\boldsymbol{\alpha}), ..., \exp(\mathbf{z}_n'\boldsymbol{\alpha}) \}$, wherein $\mathbf{Z} = (\mathbf{z}_1, ... , \mathbf{z}_n)'$ is a non-random known $(n \times r)$ matrix of observations of variables causing heteroskedasticity (with the first column consisting of ones only) and $\boldsymbol{\alpha}$ is an $(r \times 1)$ vector of unknown coefficients. In this setting, the parameters of the model are $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and both must be estimated; though only the regression parameters $\boldsymbol{\beta}$ are of interest, and it is needful to make inferences of them.

In the estimation of $\boldsymbol{\beta}$ the two-step Aitken (or feasible generalized least squares estimation) procedure is usually employed (cf. e. g. Judge, 1985, p. 422, Judge 1998, p. 352), under which the parameters $\boldsymbol{\alpha}$ are first estimated and the estimate of $\boldsymbol{\beta}$ is constructed as if the covariance matrix were known. For the purpose of inferences about $\boldsymbol{\beta}$ (and its linear combinations as of $\mathbf{L}'\boldsymbol{\beta}$ with $\mathbf{L}'$ being a reasonable $(l \times k)$ fixed matrix), conventionally, a confidence region in the form a Wald-type pivot is constructed, based upon an approximate F distribution with unadjusted degrees of freedom. The employment of an approximate F distribution is justified asymptotically, however, this self-assurance cannot be guaranteed in small samples. Kenward and Roger (1997, 2009) proposed for small-sample inference a scaled Wald-type pivot statistics with an approximate F distribution with adjusted degrees of freedom.

The aim of the article is to compare the small-sample performance of the two Kenward-Roger methods with the traditional large-sample method, when in the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ the maximum likelihood estimation is employed. Several simulation designs of the general Gaussian linear regression model under the assumption of multiplicative heteroskedasticity are set-up, albeit this assumption may not be valid.

2. The simulation study

The simulation study focuses on the performance of the Kenward-Roger methods in situations when the variable causing heteroskedasticity is identified and heteroskedastic variance components are estimated in the multiplicative form whilst the true form of heteroskedasticity may be (i.) additive, (ii.) mixed, (iii.) multiplicative, or (iv.) the variance components are of such a nature that they are unrelated to the variable identified as a cause of heteroskedasticity. Since in case (iv.) variance components behave indifferently and random towards the identified variable, this is operatively designated as the case of "indefinite" heteroskedasticity.

In all of the five general designs a simple linear regression model is explored

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{for} \quad i \in \{1,\ldots,n\}, \tag{1}$$

in which Y_i is the explained variable, x_i is the explanatory variable, β_0 and β_1 are the parameters to be estimated, and the disturbance ε_i is distributed as $N(0, \sigma^2_i(z_i))$. The dispersion of the disturbance depends on the known variable causing heteroskedasticity, which may be either (most frequently) the explanatory variable x_i itself or (seldom) a different variable $z_i \neq x_i$.

For each design the true form of heteroskedasticity appears in the functional form of $(1, z_i)'$ and a nuisance vector $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)'$. The additive model of heteroskedasticity runs as

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_i \qquad \text{for} \quad i \in \{1,\ldots,n\}, \tag{add}$$

(with $\boldsymbol{\alpha}$ such that disturbance variances σ^2_i are positive), the mixed models reads

$$\sigma_i^2 = (\alpha_0 + \alpha_1 z_i)^2 \qquad \text{for} \quad i \in \{1,\ldots,n\}, \tag{mix}$$

(with no specific restrictions laid upon $\boldsymbol{\alpha}$), and the multiplicative model respects the formula

$$\sigma_i^2 = \exp(\alpha_0 + \alpha_1 z_i) \qquad \text{for} \quad i \in \{1,\ldots,n\}, \tag{mul}$$

(again bearing no specific restrictions in respect to $\boldsymbol{\alpha}$). Eventually, a model of „indefinite" heteroskedasticity is considered, in which the disturbance components takes random values within a band around a constant, and this is determined

$$\sigma_i^2 = u_i \times \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{i=n}(\alpha_0 + \alpha_1 z_i) + \frac{1}{n}\sum_{i=1}^{i=n}\exp(\alpha_0 + \alpha_1 z_i)\right] \qquad \text{for} \quad i \in \{1,\ldots,n\}, \tag{ind}$$

where the random component u_i is distributed uniformly at some reasonable interval (ϕ_0, ϕ_1). This model applies in situations when the variable / set of variables causing heteroskedasticity is / are incorrectly specified and the variances of disturbances are not related to the identified variable / set of variables. As a consequence, with respect to the variable / variables causing heteroskedasticity they are determined randomly.

Under the assumption of multiplicative heteroskedasticity in model (1), there are several ways to the construction of a confidence region for $\boldsymbol{\beta}$ and may be structured according to the method of estimating $\boldsymbol{\alpha}$, the method of estimating $\boldsymbol{\beta}$ as well as the to the method of constructing the confidence region itself.

2.1 The estimation of β

The correct estimation of $\boldsymbol{\beta}$ requires the knowledge of $\boldsymbol{\alpha}$ (which stipulates the knowledge of $\boldsymbol{\Sigma}(\boldsymbol{\alpha})$). Should $\boldsymbol{\alpha}$ be such that the model were homoskedastic (i. e. with all but the first element zero), it would be appropriate to employ the ordinary least squares estimator

$$\mathbf{b}_{OLS} = \left(\sum_{i=1}^{i=n} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{i=n} \mathbf{x}_i Y_i \,, \tag{2}$$

should $\boldsymbol{\alpha}$ be known and model were heteroskedastic, the appropriate estimator would be that of the generalized least squares. In practical applications, when $\boldsymbol{\alpha}$ is not known and the model deemed as heteroskedastic, the two-step Aitken procedure is employed: first, the estimate of $\boldsymbol{\alpha}$ is procured and, second, the estimated generalized least squares estimator is used

$$\mathbf{b}_{EGLS}(\mathbf{a}) = \left(\sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \mathbf{x}_i Y_i \,. \tag{3}$$

This estimator coincides with the maximum likelihood estimator for $\boldsymbol{\beta}$ obtained by the method of scoring (with the respective maximum likelihood estimate of $\boldsymbol{\alpha}$ substituted for $\mathbf{a}$), as can be seen e. g. from the developments of Harvey (1976, p. 463).

2.2 The estimation of α

A rough procedure to obtain estimates of $\boldsymbol{\alpha}$ is to apply an OLS regression to model (1) and on the basis of OLS residuals $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}_{OLS}$ effect an ancillary regression $\ln e_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + \eta_i$ (in which η_i is a random disturbance term). The respective OLS estimator

$$\mathbf{a}_{OLS} = \left(\sum_{i=1}^{i=n} \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^{i=n} \mathbf{z}_i \ln e_i^2 \,, \tag{4}$$

is biased, inconsistent, and even asymptotically inconsistent (see Harvey, 1976, pp. 462-3, Judge et. al. 1985, p. 440.). For its undesirable statistical properties the OLS estimate of $\boldsymbol{\alpha}$ determined by (4) is suitable for, and utilized as, a preliminary step in iterative procedures in the construction of ML estimates with preferable properties.

Harvey (1976, p. 464) derives the first iteration of the method of scoring for the respective ML estimator based on the initial estimate $\mathbf{a}_{OLS}$ accounting for consistency bias, which is

$$\mathbf{a}_{ML} = \mathbf{a}_{OLS} + 0.2704\, \mathbf{i}_1 + 0.2807 \left(\sum_{i=1}^{i=n} \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \sum_{i=1}^{i=n} \mathbf{z}_i \exp(-\mathbf{z}_i'\mathbf{a}_{OLS})\, e_i^2 \,, \tag{5}$$

where $\mathbf{i}_1$ is an $(r \times 1)$ vector the first element of which is 1 and the remaining elements are zero. It is easy to show, by the inverse of the average information matrix, that the asymptotic covariance matrix of $\mathbf{a}_{ML}$ is

$$\operatorname{cov}(\mathbf{a}_{ML}) = 4 \left[\sum_{i=1}^{i=n} \mathbf{z}_i \mathbf{z}_i' \left(1 + \exp(-\mathbf{z}_i'\mathbf{a}_{ML})(Y_i - \mathbf{x}_i'\mathbf{b}_{ML})^2 \right) \right]^{-1} , \tag{6}$$

in which $\mathbf{b}_{ML}$ is the corresponding two-step Aitken estimator $\mathbf{b}_{EGLS}(\mathbf{a}_{ML})$.

2.3 The construction of a confidence region for β

The standard procedure for the construction of a confidence region for $\boldsymbol{\beta}$, or more precisely for its linear combinations in the form $\mathbf{L}'\boldsymbol{\beta}$, wherein $\mathbf{L}'$ is a reasonable $(l \times k)$ fixed matrix, is described in detail in Judge et al. (1988, p. 355). To this end, a commonly used approach is founded on the statistic

$$F = \frac{(\mathbf{b}_{EGLS} - \boldsymbol{\beta})' \mathbf{L} \left[\sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a})(\mathbf{L}'\mathbf{x}_i\mathbf{x}_i'\mathbf{L}) \right]^{-1} \mathbf{L}'(\mathbf{b}_{EGLS} - \boldsymbol{\beta})}{\sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a})(Y_i - \mathbf{x}_i'\mathbf{b}_{EGLS})^2} \cdot \frac{n-k}{l}, \tag{7}$$

where $\mathbf{a}$ is a reasonable estimate of $\boldsymbol{\alpha}$ and $\mathbf{b}_{EGLS}$ is the respective two-step Aitken estimator of $\boldsymbol{\beta}$. This statistic is under the hypothesis $\mathbf{L}'\mathbf{b}_{EGLS} = \mathbf{L}'\boldsymbol{\beta}$ distributed approximately as $F(l, n - k)$.

Despite the fact that the statistic given in (10) has been found relatively accurate in finite samples (cf. Judge et al. 1988, p. 355), the usage of such confidence region is justified only asymptotically. For small samples, the method of Kenward and Roger is appropriate. Kenward and Roger (1997, 2009) proposed a scaled F statistic with small-sample covariance matrix for the estimator $\mathbf{b}_{EGLS}$. The original Kenward-Roger covariance matrix of $\mathbf{b}_{EGLS}$ as proposed in 1997, for the model of multiplicative heteroskedasticity, is

$$(\operatorname{cov} \mathbf{b}_{EGLS})_{OLD} = \hat{\boldsymbol{\Phi}} + 2\hat{\boldsymbol{\Phi}} \left\{ \sum_{s=1}^{s=k} \sum_{t=1}^{t=k} \{\operatorname{cov} \mathbf{a}\}_{st} \left(\tfrac{3}{4} \hat{\mathbf{Q}}_{st} - \hat{\mathbf{P}}_s \hat{\boldsymbol{\Phi}} \hat{\mathbf{P}}_t \right) \right\} \hat{\boldsymbol{\Phi}}, \tag{8}$$

where cov $\mathbf{a}$ is an appropriate estimate of the covariance matrix of $\boldsymbol{\alpha}$ and for $s,t \in \{1, ..., k\}$

$$\hat{\boldsymbol{\Phi}}^{-1} = \sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \mathbf{x}_i \mathbf{x}_i', \qquad \hat{\mathbf{Q}}_{ij} = \sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \{\mathbf{z}_i\}_s \{\mathbf{z}_i\}_t \mathbf{x}_i \mathbf{x}_i', \qquad \hat{\mathbf{P}}_s = \sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \{\mathbf{z}_i\}_s \mathbf{x}_i \mathbf{x}_i',$$

with the computed estimate $\mathbf{a}$ inserted instead of the estimator $\mathbf{a}$.

Later Kenward and Roger in 2009 allowed for possible bias in $\mathbf{a}$, and incorporated in (8) an additional bias adjustment term so that the covariance matrix of $\mathbf{b}_{EGLS}$ in this case is estimated by

$$(\operatorname{cov} \mathbf{b}_{EGLS})_{NEW} = (\operatorname{cov} \mathbf{b}_{EGLS})_{OLD} - \frac{1}{4} \sum_{s=1}^{s=k} \sum_{t=1}^{t=k} \{\operatorname{cov} \mathbf{a}\}_{st} \upsilon_s \hat{\boldsymbol{\Phi}} \hat{\mathbf{P}}_t \hat{\boldsymbol{\Phi}}, \tag{9}$$

in which, again, cov $\mathbf{a}$ is a reasonable estimate of the covariance matrix of $\boldsymbol{\alpha}$ and the scalars υ_s are defined for $s \in \{1, ..., k\}$ as

$$\upsilon_s = \sum_{p=1}^{p=k} \sum_{r=1}^{r=k} \{\operatorname{cov} \mathbf{a}\}_{pr} \left[\sum_{i=1}^{i=n} \{\mathbf{z}_i\}_p \{\mathbf{z}_i\}_r \{\mathbf{z}_i\}_s - 2 \operatorname{Tr} \left[\left(\sum_{i=1}^{i=n} \exp(-\mathbf{z}_i'\mathbf{a}) \{\mathbf{z}_i\}_p \{\mathbf{z}_i\}_r \{\mathbf{z}_i\}_s \mathbf{x}_i \mathbf{x}_i' \right) \hat{\boldsymbol{\Phi}} \right] - \operatorname{Tr} \hat{\mathbf{Q}}_{pr} \hat{\boldsymbol{\Phi}} \hat{\mathbf{P}}_s \hat{\boldsymbol{\Phi}} \right],$$

with the substituted computed estimate $\mathbf{a}$ and the estimate cov $\mathbf{a}$.

In the construction of confidence regions for $\mathbf{L}'\boldsymbol{\beta}$ Kenward and Roger (1997, 2009) advocated the employment of a Wald-type pivot of the form

$$F^* = \frac{\lambda}{l} (\mathbf{b}_{EGLS} - \boldsymbol{\beta})' \mathbf{L} (\mathbf{L}' \operatorname{cov} \mathbf{b}_{EGLS} \mathbf{L})^{-1} \mathbf{L}' (\mathbf{b}_{EGLS} - \boldsymbol{\beta}), \tag{10}$$

with their proposed covariance matrix cov $\mathbf{b}_{EGLS}$, which under the hypothesis $\mathbf{L}'\mathbf{b}_{EGLS} = \mathbf{L}'\boldsymbol{\beta}$ follows an approximate $F(l, m)$ distribution, whereas

$$m = 4 + \frac{2l+4}{l \frac{V}{E^2} - 2}, \qquad \lambda = \frac{m}{E(m-2)}, \qquad E = 1 + \frac{1}{l} A_2, \qquad V = \frac{1}{l} \left(2 + \frac{1}{l} (A_1 + 6A_2) \right),$$

$$A_1 = \sum_{s=1}^{s=k} \sum_{t=1}^{t=k} \{\operatorname{cov} \mathbf{a}\}_{ij} \operatorname{Tr} \left\{ \mathbf{L} (\mathbf{L}' \operatorname{cov} \mathbf{b}_{EGLS} \mathbf{L})^{-1} \mathbf{L}' \hat{\boldsymbol{\Phi}} \mathbf{P}_s \hat{\boldsymbol{\Phi}} \right\} \operatorname{Tr} \left\{ \mathbf{L} (\mathbf{L}' \operatorname{cov} \mathbf{b}_{EGLS} \mathbf{L})^{-1} \mathbf{L}' \hat{\boldsymbol{\Phi}} \mathbf{P}_t \hat{\boldsymbol{\Phi}} \right\},$$

$$A_2 = \sum_{s=1}^{s=k} \sum_{t=1}^{t=k} \{\operatorname{cov} \mathbf{a}\}_{ij} \operatorname{Tr} \left\{ \mathbf{L} (\mathbf{L}' \operatorname{cov} \mathbf{b}_{EGLS} \mathbf{L})^{-1} \mathbf{L}' \hat{\boldsymbol{\Phi}} \mathbf{P}_s \hat{\boldsymbol{\Phi}} \mathbf{L} (\mathbf{L}' \operatorname{cov} \mathbf{b}_{EGLS} \mathbf{L})^{-1} \mathbf{L}' \hat{\boldsymbol{\Phi}} \mathbf{P}_t \hat{\boldsymbol{\Phi}} \right\}.$$

2.4 The simulation framework

In the simulation study, the aforementioned ML-based estimator for $\boldsymbol{\alpha}$ was considered. Having at hand the estimate of $\boldsymbol{\alpha}$, the two-step Aitken estimate $\mathbf{b}_{EGLS}$ of $\boldsymbol{\beta}$ was computed, and a confidence

region for $\boldsymbol{\beta}$ (viz., with $\mathbf{L}'$ set to an identity matrix) was set-up according to all of the three methods for the construction of confidence regions: the standard confidence region as given in (7), the Kenward-Roger confidence region of (10) with the "old" covariance matrix of $\mathbf{b}_{EGLS}$ (8) and the Kenward-Roger confidence region of (10) with the "new" covariance matrix of $\mathbf{b}_{EGLS}$ (9). As a consequence, the total count of 3 confidence regions was examined.

Five designs were considered, each in 4 modifications according to the parameters of the sampling distribution and the set values of $\boldsymbol{\alpha}$. **Design A** comprised x_i's and z_i's in (1) sampled from a uniform population, such that for sub-design Axx $z_i = x_i$ and for sub-design Axz $z_i \neq x_i$. **Design B** was formed of x_i's and z_i's drawn from a Gaussian population, such that for sub-design Bxx $z_i = x_i$ and for sub-design Bxz $z_i \neq x_i$. **Design C** (Cxz) consisted in x_i's distributed uniformly and z_i's Gaussian and **design D** (Dxz), reversely, was based upon x_i's Gaussian and z_i's uniform. The last **design E** was a drawing from a lognormal distribution, again such that for sub-design Exx $z_i = x_i$ and for sub-design Exz $z_i \neq x_i$. Random values of x_i's and z_i's were rounded to two decimals, the values for the explanatory variable y_i were simulated according to (1), given the model of heteroskedasticity, and were not rounded. The computations ran first for 10 observations, to which 5 more were added, and the sample of 15 observations was created. To this sample 5 more observations were appended etc.; the simulations were based on samples counting 10, 15, 20, 30 and 50 observations, which were fixed for each evaluation out of 10 000 simulations.

For each design and its modifications, the parameters of the considered distributions and the vector $\boldsymbol{\alpha}$ were determined so that in some designs and their modifications an occurrence of negative values was permitted (for uniform and Gaussian cases) as well as negative values in $\boldsymbol{\alpha}$ were accounted for. The presence of negative values in economic variables is not hypothetical and is not atypical, frequent examples include net assets value, net income or net interests which take oftentimes negative values. The vector of regression parameters was $\boldsymbol{\beta} = (10, 6)'$ in each simulation design and its modifications. As stated in Glejser (1969, p. 319), the choice of $\boldsymbol{\beta}$ is irrelevant.

However, room here does not suffice to present the full results of the simulation study, therefore, for an illustration, only the results of selected three cases are shown in Table. The details and results of the remaining 29 cases are available from the author on request. Under the term coverage the proportion of cases in which simulated data were covered by the respective confidence region statistic at 0.95 confidence level is understood.

Table. The results of the selected simulation designs (Source: the author.)

The coverage results of the selected simulation designs													
Confidence region of $\boldsymbol{\beta}$		1997 KR	2009 KR	Standard method	1997 KR	2009 KR	Standard method	1997 KR	2009 KR	Standard method	1997 KR	2009 KR	Standard method
Design Axx1	Obs.	TRUE ADDITIVE MODEL			TRUE MULTIPLICATIVE MODEL			TRUE MIXED MODEL			TRUE "INDEFINITE" MODEL		
xi = zi U (0, 100) $\boldsymbol{\alpha}$ = (5, 0.25)' $\boldsymbol{\beta}$ = (10, 6)'	10	0.8322	0.8698	0.8615	0.7354	0.7512	0.8613	0.8522	0.8776	0.8654	0.8282	0.8738	0.8681
	15	0.8947	0.917	0.8968	0.9983	0.9984	0.9907	0.9155	0.9311	0.91	0.8802	0.9063	0.8848
	20	0.9254	0.9373	0.9232	0.9971	0.9971	0.9894	0.944	0.9537	0.9347	0.9217	0.9362	0.9193
	30	0.9445	0.9511	0.9404	0.9973	0.9982	0.9788	0.9543	0.9626	0.9494	0.9378	0.9478	0.9339
	50	0.9523	0.957	0.9482	0.9984	0.9988	0.98	0.956	0.9612	0.9507	0.9413	0.9474	0.9362
Design Bxx1	Obs.	TRUE ADDITIVE MODEL			TRUE MULTIPLICATIVE MODEL			TRUE MIXED MODEL			TRUE "INDEFINITE" MODEL		
xi = zi N (2, 2.5) $\boldsymbol{\alpha}$ = (2, 0.03)' $\boldsymbol{\beta}$ = (10, 6)'	10	0.8533	0.8795	0.8745	0.8394	0.8702	0.8667	0.8402	0.8698	0.8634	0.8483	0.8779	0.872
	15	0.8864	0.9056	0.8886	0.8883	0.9083	0.8863	0.8895	0.908	0.8882	0.8924	0.9098	0.8923
	20	0.9077	0.9217	0.8982	0.9097	0.9228	0.9003	0.9095	0.9229	0.8985	0.9107	0.9243	0.9015
	30	0.9288	0.9382	0.9195	0.9282	0.9371	0.9156	0.9308	0.9406	0.9206	0.9285	0.9372	0.9158
	50	0.9398	0.9455	0.9323	0.9418	0.9469	0.9336	0.9438	0.9484	0.9357	0.9432	0.9482	0.9343
DesignExx1	Obs.	TRUE ADDITIVE MODEL			TRUE MULTIPLICATIVE MODEL			TRUE MIXED MODEL			TRUE "INDEFINITE" MODEL		
xi = zi Λ (2, 2.5) $\boldsymbol{\alpha}$ = (7, -0.07)' $\boldsymbol{\beta}$ = (10, 6)'	10	0.9345	0.9515	0.9471	0.8242	0.8605	0.79	0.9338	0.9509	0.9339	0.8955	0.9173	0.9436
	15	0.9572	0.9707	0.9551	0.8831	0.905	0.8073	0.9642	0.9742	0.9468	0.9139	0.9341	0.9456
	20	0.8956	0.9301	0.9101	0.9613	0.9729	0.9518	0.9524	0.9716	0.9522	0.8461	0.8825	0.85
	30	0.9197	0.9416	0.9125	0.9515	0.9637	0.9389	0.9633	0.9726	0.9524	0.892	0.9145	0.8845
	50	0.9812	0.9842	0.9719	0.9757	0.9803	0.9588	0.9877	0.9896	0.9762	0.9302	0.9411	0.9213

Inspecting the results when the true form of heteroskedasticity is additive, it is found that the coverage with all the six methods for the construction of a confidence region is satisfied from 30 observations, in some designs from 50 observations, and it seems that both Kenward-Roger methods with the "new" covariance matrix have a higher coverage, in some designs 0.96 – 0.98 at 50

observations. There cannot be seen influence of the sampling distribution or of negative values upon the performance.

When the true heteroskedasticity is multiplicative, i. e. when the model of heteroskedasticity is correctly specified, in general, all methods give a good coverage of approximately 0.95 or higher at 30 or 50 observations. Having stated this, it must be said that with some designs (any x_i's and any z_i's both uniform, Gaussian negative x_i's and uniform positive z_i's, Gaussian x_i's and uniform z_i's both negative, log-normal x_i's and z_i's with positive parameters) the coverage is almost 1 with the Kenward-Roger methods, starting at 10 or 20 observations. Otherwise, on the whole, it seems that the 1997 Kenward-Roger method exhibits a better coverage, though frequently above 0.95 at 50 observations.

In the cases when the true model heteroskedasticity is mixed, in the design of x_i's and z_i's both uniform with negative values in $\boldsymbol{\alpha}$, the coverage of 0.95 is mostly attained at 15 or 20 observations and increases with rising number of observations so that at 50 observations, all the methods report coverage approx. 0.96 or much higher (almost 1). A similar situation is with x_i's Gaussian and z_i's uniform with negative values in $\boldsymbol{\alpha}$ or variables (save the case when both $\boldsymbol{\alpha}$ and variables contain negative values), and with x_i's and z_i's both log-normal with positive values in $\boldsymbol{\alpha}$. Again, the 1997 Kenward-Roger method shows a better coverage with an upward bias at 30 or 50 observations.

Going over the cases of the true "indefinite" heteroskedasticity, in all the designs the coverage rate with all the methods is slightly below 0.95 or approximately 0.95 at 50 observations. The Kenward-Roger method with the "new" covariance matrix appears to give a good (exact) coverage of 0.95 at 50 observations.

In addition to these findings, there may be observed a higher coverage in situations when either the variable x_i or z_i or the parameter $\boldsymbol{\alpha}$ contain negative values, with bias upwards.

3. The conclusion

In constructing a confidence region for a linear combination of the regression parameters of the general Gaussian linear regression model with the multiplicative form of heteroskedasticity several inference methods are available, and they are structured according to (i.) the estimation procedure they employ to arrive at the estimates of the nuisance parameter $\boldsymbol{\alpha}$ and (ii.) the method for the set-up of a confidence region for $\boldsymbol{\beta}$ (and its linear combinations) itself. In the article, from amongst estimators of $\boldsymbol{\alpha}$, the ML estimator and the REML estimator are considered, and in the very construction of a confidence region for $\boldsymbol{\beta}$, the conventional approach and the two Kenward-Roger methods are applied. In the Monte Carlo simulation study, a Gaussian linear regression model with multiplicatively heteroskedastic disturbances with one explanatory variable is postulated and the performance of the three methods is evaluated in 5 basic designs with 4 modifications each as to the sampling distribution and the (true unknown) values of $\boldsymbol{\alpha}$ in situations when the model of multiplicative heteroskedasticity is identified irrespective of what true underlying heteroskedasticity is. It is found that all the methods considered are quite robust to the correct specification of the true form of heteroskedasticity, which might imply that the correct identification of the heteroskedasticity model may not be of great importance as the model of multiplicative heteroskedasticity is found sufficiently descriptive. The desired coverage rate of the confidence region with each method considered is found attained at the sample of 50 observations (and in most cases even fewer), however, there may be expected bias upwards in the coverage rate, especially in cases when the explanatory variable or the variable causing heteroskedasticity contain negative values. Furthermore, the Kenward-Roger procedures in the construction of the confidence region may be superior to the conventional procedure, and can be recommended for the samples of below 50 observations for its higher coverage.

4. The literature

[1] GLEJSER, H. 1969. A new test for heteroskedasticity. In: Journal of the American Statistical Association, 1969, vol. 64, iss. 325, pp. 316-323. ISSN 0162-1459.

[2] HARVEY, A. C. 1976. Estimating regression models with multiplicative heteroskedasticity. In: Econometrica, 1976, vol. 44, iss. 3, pp. 461-465. ISSN 0012-9682.

[3] JUDGE, G. G. et al. 1985 The theory and practice of econometrics. 2nd ed. New York: Wiley, 1985. 1019 p. ISBN 978-0-471-89530-5.

[4] JUDGE, G. G. et al. 1988. Introduction to the theory and practice of econometrics. 2nd ed. New York: Wiley, 1988. 1024 p. ISBN 0-471-62414-4.

[5] KENWARD, M. G., ROGER, J. H. 1997. Small sample inference for fixed effects from restricted maximum likelihood. In: Biometrics, 1997, vol. 53, iss. 3, pp. 983-997. ISSN 0006-341X.

[6] KENWARD, M. G., ROGER, J. H. 2009. An improved approximation to the precision of fixed effects from restricted maximum likelihood. In: Computational Statistics and Data Analysis, 2009, vol. 53, iss. 7, pp. 2583-2595. ISSN 0167-9473.

[7] MEYER, K. 1997. An average information' restricted maximum likelihood algorithm for estimating reduced rank genetic covariance matrices or covariance functions for animal models with equal design matrices. In: Genetics Selection Evolution, 1997, vol. 29, pp. 97-116. ISSN 0999-193X.

The author's address

Martin Boďa, Ing. et Bc.
Univerzita Mateja Bela v Banskej Bystrici
Fakulta prírodných vied, Katedra matematiky
Tajovského 40, 974 01 Banská Bystrica
Ekonomická fakulta, Katedra kvantitatívnych metód a informačných systémov
Tajovského 10, 975 90 Banská Bystrica
martin.boda@umb.sk

Aktuálne otázky na trhu s bývaním na Slovensku
Current issues in the housing market in Slovakia

Mikuláš Cár

Abstract: Information needs on the Slovak housing market are from the perspective of its players very differential and currently are not satisfactorily resolved. Even a relatively briefly functioning housing market in Slovakia confirmed its cyclical development. The young Slovak housing market has gradually stabilised in line with economic fundamentals. Current development of selected indicators on the demand side of Slovak housing market indicates a certain momentum to its gradual stabilization.

Key words: housing market, data sources, residential property prices, supply and demand factors.

Kľúčové slová: trh s bývaním, zdroje údajov, ceny nehnuteľností na bývanie, ponukové a dopytové faktory.

1 Úvod

Trh s bývaním je významným segmentom realitného trhu a na Slovensku sa začal kreovať v podstate len na prelome milénia. Už v polovici deväťdesiatych rokov však boli urobené prvé kroky, ktoré prispeli k postupnému vytváraniu podmienok pre vznik trhu s bývaním aj na Slovensku. Dôležitým krokom v tomto smere bola novelizácia legislatívy, ktorá upravila možnosti prevodu bytov z vlastníctva bytových družstiev a obcí do osobného vlastníctva a vytvoril sa tak predpoklad na postupné uplatňovanie trhového princípu pri nakladaní s bytmi[1]. Významným príspevkom ku kreovaniu slovenského realitného trhu bol príchod hypotekárneho bankovníctva, ktoré postupne výrazne rozširovalo úverové možnosti pre záujemcov o bývanie[2].

[1] Realizácia na základe zmluvy o prevode vlastníctva bytu a zriadení záložného práva, uzatváranej podľa §5, ods. 1 Zákona NR SR č. 182/1993 Z. z. v znení Zákona NR SR č. 151/1995 Z. z. o vlastníctve bytov a nebytových priestorov. Takto bolo možné získať do osobného vlastníctva obecné byty za výhodných finančných podmienok s tým, že nadobudateľ sa zaviazal, že počas dohodnutého obdobia (v podmienkach Bratislavy napr. počas 10 rokov) od podpísania zmluvy, neprevedie vlastníctvo predmetného bytu na inú osobu ako na manžela, deti, vnukov alebo rodičov. Na predmetný prevedený byt bolo na ten čas zriadené záložné právo v katastri nehnuteľností a toto záložné právo zaniklo buď uplynutím času, alebo vyplatením zrážky, resp. zľavy z ceny bytu, ktorá bola poskytnutá pri prevode vlastníctva daného bytu.

[2] Prijatá novela Zákona NR SR č. 21/1992 Z. z. o bankách v podobe zákona č. 58/1996 Z. z., ktorá vstúpila do platnosti 1.3.1996. K výraznejšiemu presadeniu sa hypotekárneho úverovania došlo až po 1.4.1999 (účinnosť novely zákona č. 286/1992 Z. z. o daniach z príjmov v znení neskorších predpisov), ktorá prispela k priamej i nepriamej štátnej podpore hypotekárneho bankovníctva prostredníctvom oslobodenia úrokových výnosov z hypotekárnych záložných listov od dane z príjmu. Táto skutočnosť sa stala zaujímavá pre potenciálnych investorov do hypotekárnych záložných listov (v zmysle zákona o bankách (Zákon NR SR č. 483/2001 Z. z.) slovenské hypotekárne banky získavajú zdroje na financovanie hypotekárnych úverov najmä prostredníctvom predaja hypotekárnych záložných listov). Efektívna úroková sadzba pre fyzické osoby - občanov postupne klesla na úroveň okolo 7,25 - 7,5 %, čím sa značne priblížila úrovni úrokových sadzieb stavebných sporiteľní pri stavebných úveroch a vo vzťahu k medziúverom stavebných sporiteľní začala byť úroková sadzba hypotekárnych úverov dokonca výhodnejšia.

Rozhodujúcim momentom pre vznik trhu s bývaním na Slovensku však bola očividne rastúca výkonnosť slovenskej ekonomiky po roku 2000, čo sa prejavilo aj na zlepšovaní sa príjmovej situácie obyvateľstva a tiež na vytváraní pozitívnych očakávaní do budúcnosti. S tým súvisela aj pomerne výrazná ochota zadlžiť sa pri zabezpečovaní si vlastného bývania mladými domácnosťami, resp. pri zlepšovaní komfortu bývania strednej a staršej generácie. V tom čase sa naďalej zlepšovali podmienky pre získanie úverových zdrojov na obstaranie si bývania a objavili sa aj záujemcovia o nákup domov a bytov na účely zhodnotenia investície. Ak k tomu pridáme vstup Slovenska do EÚ v roku 2004 a s tým spojený príchod záujemcov o byty aj spoza hraníc, výsledkom bol výrazne rastúci dopyt po domoch a bytoch, za ktorým zaostávala reálna ponuka. Prejavilo sa to na prudkom raste cien nehnuteľností na bývanie, ktorý sa skončil v druhom štvrťroku 2008 aj v dôsledku prenikania turbulencií zo zahraničných realitných trhov.

Doterajší vývoj na relatívne mladom slovenskom trhu s bývaním bol pomerne turbulentný a súvisel do značnej miery aj s prebiehajúcimi konvergenčnými procesmi Slovenska smerom k EÚ. Súčasný slovenský realitný trh už možno považovať za relatívne štandardizovaný, postupne sa však ďalej kryštalizuje a stabilizuje v súlade s vývojom ekonomických fundamentov.

2 Údajová základňa o slovenskom trhu s bývaním

Vývoj na realitných trhoch vo svete v posledných rokoch priniesol celý rad poznatkov a skúseností, že nehnuteľnosti môžu nielen pozitívne prispievať k zlepšovaniu výkonnosti celkovej ekonomiky, ale môžu byť aj zdrojom vážnych porúch v transakčnom mechanizme. To vyvoláva potrebu mať dostatok informácií o jednotlivých segmentoch a rôznych stránkach realitného trhu. Záujem o informácie ohľadom realitného trhu rastie u všetkých jeho aktérov. Je pochopiteľná rôznorodosť ich záujmov, čo sa prejavuje aj na požiadavkách jednotlivých subjektov realitného trhu na obsah i rozsah údajov.

V podmienkach Slovenska sa vytvára informačná základňa o jednotlivých segmentoch realitného trhu relatívne izolovane, hlavne podľa možností jednotlivých subjektov. Obecne zatiaľ absentuje jednotiaca metodika a systém vytvárania takej databázy údajov o realitnom trhu, ktorá by poskytovala potrebné informácie všetkým zainteresovaným subjektom, a aby údaje vo vytvorenej databáze boli aj medzinárodne porovnateľné. Aktuálny stav je do značnej miery objektívne daný jednak krátkym časom fungovania slovenského realitného trhu, ale aj zložitosťou danej problematiky. Objavujú sa však už aj určité náznaky spoločného hľadania riešení existujúceho stavu.

K parciálnym i čiastočne systematizovaným údajom o slovenskom trhu s bývaním sa možno dostať prostredníctvom viacerých internetových realitných portálov. Obsahujú celý rad aktuálnych a užitočných informácií pre záujemcov o predajoch, nákupoch, prenájmoch napr. nielen bytov, domov a novostavieb, ale aj podnikateľských priestorov, rekreačných objektov, pozemkov, garáží a pod. a to aj v určitom regionálnom členení. Poskytujú aj určité štatistiky o vývoji priemerných cien vybraných druhov nehnuteľností na bývanie na týždňovej báze zhruba za posledný polrok a možno sa dopracovať aj k určitým časovým radom vybraných ukazovateľov za dlhšie časové obdobie.[1] Určitým kvalitatívnym posunom v smere k vytvoreniu bohato štruktúrovanej databázy nielen o cenách domov a bytov, ale aj ďalších segmentov na realitnom trhu je internetový portál Cenová mapa nehnuteľností (www.cmn.sk),

[1] Určitý prehľad realitných portálov, prevádzkovaných na Slovensku a v Českej republike možno nájsť napr. na adrese: http://www.zoznam.sk/katalog/Spravodajstvo-informacie/prevadzkovatelia-portalov/prevadzkovatelia-realitnych-portalov/.

ktorý je softvérovou nadstavbou na údajoch, ktoré sú zhromažďované prostredníctvom realitných kancelárií.

Užitočné štatistiky o rôznych aspektoch ponukovej a dopytovej stránky realitného trhu sú dostupné v prierezových národných štatistikách a v jednotlivých národných rezortných štatistikách. Z globálneho pohľadu sú dôležité informácie o možných vonkajších vplyvoch na domáci realitný trh ako aj základné informácie o doterajšom vývoji a aktuálnej kondícii domácej reálnej ekonomiky, od ktorých sa ďalej odvíjajú jednotlivé ponukové a dopytové faktory na realitnom trhu.

Z ponukovej stránky môžu byť zaujímavé napr. informácie o:

- počtoch stavebných povolení,
- počtoch začatých, dokončených a rozostavaných bytov,
- údaje o objeme investícií smerujúcich do jednotlivých realitných segmentov,
- údaje o objeme produkcie v stavebnej výrobe spojenej s výstavbou bytových budov a pod.

Z dopytovej stránky môžu byť zaujímavé napr. informácie o:

- vekovom zložení obyvateľstva,
- aktuálnych príjmových možnostiach a perspektívnych očakávaniach obyvateľstva,
- úsporách a úveroch obyvateľstva,
- úverových podmienkach a dostupnosti úverov v relevantných úverových inštitúciách a pod.

Z relácie medzi dopytom a ponukou na trhu s bývaním rezultujú ceny nehnuteľností na bývanie. V záujme poznania aktuálneho stavu, možnosti regionálneho porovnávania a posúdenia vplyvu rôznych faktorov na vývoj cien domov a bytov je potrebné mať k dispozícii funkčnú a podľa možnosti jednotne vymedzenú databázu údajov. Vzhľadom na krátky čas trvania reálneho trhu s bývaním na Slovensku, existuje databáza priemerných cien nehnuteľností na bývanie na ročnej báze od roku 2002 a na štvrťročnej báze od začiatku roku 2005 v réžii Národnej banky Slovenska (NBS).

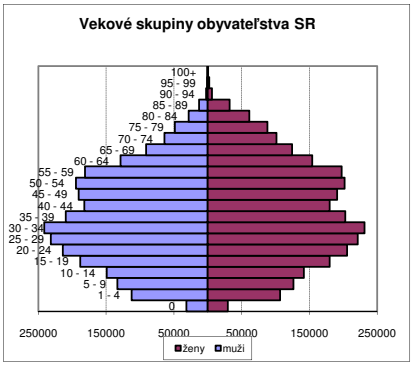
Štvrťročné zisťovanie priemerných cien nehnuteľností na bývanie je realizované na základe spracovanej metodiky a používajú sa údaje Národnej asociácie realitných kancelárií Slovenska (NARKS), ktoré NBS získava na zmluvnom základe. Zhruba mesiac po referenčnom štvrťroku sú na web stránke NBS pravidelne zverejňované údaje o priemerných cenách jednotlivých druhov bytov a domov aj v regionálnom členení podľa krajov.

Rastúce požiadavky subjektov realitného trhu na rozsah i kvalitu údajov vyvolávajú potrebu hľadať možnosti využitia ďalších existujúcich i potenciálnych zdrojov údajov. Ako o významných potenciálnych zdrojoch údajov ohľadom zisťovania cien domov a bytov sa často hovorí o katastri nehnuteľností a o úverovom registri. Významné aktivity v tomto smere vyvíja aj Štatistický úrad SR, ktorý v rámci pilotného projektu Eurostatu pripravuje na pravidelné publikovanie Residential Property Price Indices podľa jednotnej celoeurópskej metodiky[1].

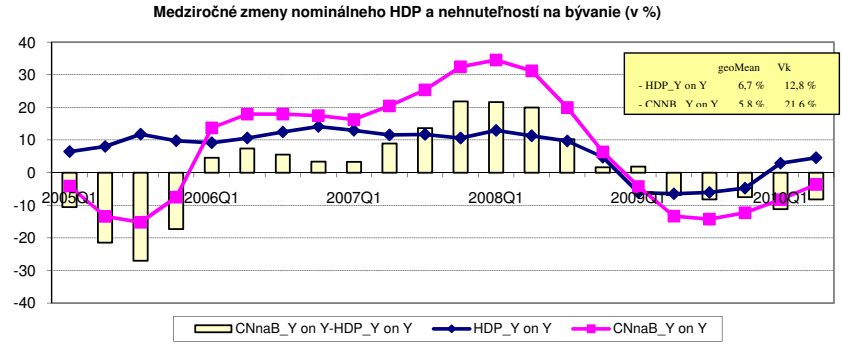
[1] Bližšie pozri Handbook on Residential Property Price Indices. Eurostat, May 2010. (http://epp.eurostat.ec.europa.eu/portal/page/portal/hicp/documents/Tab/Tab/rppi_handbook%20_combinedv1_0.pdf)

3 Vybrané impulzy determinujúce aktuálny vývoj na slovenskom trhu s bývaním

Na základe dostupných zdrojov údajov možno dospieť k poznaniu, že ekonomická recesia spôsobila v podmienkach Slovenska postupné spomaľovanie až zastavenie dopytu po bývaní. Súčasne však ešte dobiehala realizácia rozbehnutých rezidenčných projektov, čo spôsobilo stav, že aktuálne ponuka domov a bytov prevyšuje dopyt. Z vývoja vybraných indikátorov na dopytovej stránke slovenského trhu s bývaním sa dajú vypozorovať určité impulzy pre jeho ďalší vývoj.



Bývanie je permanentne aktuálnou témou v rámci životného cyklu populácie. Určité demografické aspekty životného cyklu dokážu významným spôsobom ovplyvniť dopytovú stránku na trhu s bývaním Snaha zabezpečiť si vlastné bývanie mladej generácie je jednou z hlavných súčastí dopytovej stránky realitného trhu. V tomto smere je potrebné venovať zvýšenú pozornosť tzv. populačným vlnám, ktoré majú v určitom období tendenciu výraznejším spôsobom zvýšiť dopyt po bývaní, ako je to obvykle. V podmienkach Slovenska sa zvyklo v poslednom období hovoriť o populačnej vlne v súvislosti s generáciou narodenou v sedemdesiatych a osemdesiatych rokoch minulého storočia, ktorá v súčasnosti dosahuje vek spojený so zakladaním vlastných rodín, prípadne osamostatňovaním sa od rodičov. Aktuálne vyše 925 tisícová skupina obyvateľov vo veku od 25 do 34 rokov predstavuje významnú masu, ktorej značná časť predstavuje záujemcov o uspokojenie vlastného bývania. Dopytovú stránku zvyšuje aj rastúca rozvodovosť a rastúci počet jednočlenných domácností. Nezanedbateľný vplyv na rast dopytu po bývaní má aj snaha strednej a staršej generácie o zmenu a skvalitnenie existujúceho bývania. Znamená to, že súčasné vekové zloženie obyvateľstva stále predstavuje významný dopytový faktor na slovenskom trhu s bývaním.



Zdroj: ŠÚ SR, NARKS, výpočty a graf NBS.

Už aj relatívne krátky čas fungovania realitného trhu na Slovensku potvrdil jeho cyklický vývoj. Vývoj priemerných cien nehnuteľností na bývanie pomerne výrazne reaguje na vývoj reálnej ekonomiky, ale s určitým časovým oneskorením. Od začiatku roku 2005 rástol na Slovensku HDP na medziročnej báze v priemere mierne výraznejšie (o 6,7 %) ako

priemerné ceny nehnuteľností na bývanie (o 5,8 %), ale medziročný vývoj priemerných cien domov a bytov bol výrazne variabilnejší ako vývoj HDP (variačný koeficient v prípade medziročného vývoja priemerných cien nehnuteľností dosiahol hodnotu takmer 22 %, kým v prípade medziročného vývoja HDP približne 13 %). Medziročný rast HDP je od začiatku roku 2010 sprevádzaný postupným znižovaním medziročného poklesu priemerných cien nehnuteľností na bývanie, ktorý sa vo 4. štvrťroku 2010 pravdepodobne zmení po dvoch rokoch na mierny medziročný rast. Za celý rok 2010 sa celkovo očakáva medziročný pokles priemerných cien domov a bytov asi o tri percentá v porovnaní s rokom 2009, kedy bol zaznamenaný medziročný pokles o viac ako 11 percent.

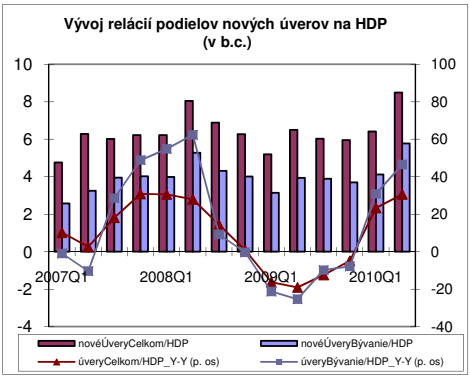
Podľa výpočtov NBS z údajov NARKS vyplýva, že priemerná cena metra štvorcového nehnuteľností na bývanie dosiahla na Slovensku v 3. štvrťroku 2010 úhrnnú hodnotu 1304 €/m2, čo predstavuje zvýšenie o 11 euro v porovnaní s 2. štvrťrokom 2010. Po viac ako dvoch rokoch došlo k obnoveniu medzištvrťročného rastu o 0,9 % a spomalil sa medziročný pokles v porovnaní s predchádzajúcim štvrťrokom o 2,3 percentuálneho bodu na -1,4 %.

Celoslovenský priemer ovplyvňuje predovšetkým vývoj priemerných cien nehnuteľností na bývanie v Bratislavskom kraji, ktorý má najväčší podiel na slovenskom trhu s bývaním. Väčšiu váhu má obchodovanie s bytmi ako s domami. Výrazný rast cien bytov po roku 2005 spôsobil, že aktuálne predstavuje priemerná cena metra štvorcového domu necelých 90 % priemernej ceny metra štvorcového bytu. Priemerné ceny bytov sú na Slovensku už dlhodobejšie nad úhrnnou cenou metra štvorcového nehnuteľností na bývanie, t. j. nad spoločnou priemernou cenou metra štvorcového domov a bytov (posledné dva roky takmer o 3 %).

V polovici slovenských krajov (BA, KE, TT a NR) došlo v 3. štvrťroku 2010 k zvýšeniu priemerných cien nehnuteľností na bývanie, pričom oproti predchádzajúcemu štvrťroku najvýraznejšie vzrástli priemerné ceny metra štvorcového obytnej plochy nehnuteľností na bývanie v Košickom kraji (3,3 %) a najvýraznejšie medzištvrťročne poklesli v Trenčianskom kraji (-2,6 %). Vývoj priemerných cien nehnuteľností na bývanie bol v posledných dvoch štvrťrokoch relatívne variabilný vo väčšine slovenských krajov (okrem Bratislavského kraja a Trenčianskeho kraja).

Na medziročnej báze došlo v 3. štvrťroku 2010 v porovnaní s predchádzajúcim štvrťrokom vo všetkých slovenských krajoch buď k zrýchleniu medziročného rastu (BB), prechodu z poklesu na rast (KE, BA) alebo k spomaleniu medziročného poklesu priemerných cien nehnuteľností na bývanie (TT, NR, TN, ZA a PO).

Dôležitým impulzom, ktorý významnou mierou zdynamizoval slovenský realitný trh v poslednom desaťročí bolo zlepšenie dostupnosti úverov na bývanie. Prosperujúca ekonomika vytvorila predpoklad pre zlepšenú príjmovú situáciu a pozitívne očakávania obyvateľstva, s čím súvisí ochota domácností sa aj zadlžiť v snahe obstarať si, alebo skvalitniť vlastné bývanie. Novo poskytnuté úvery na bývanie predstavujú zhruba dve tretiny celkového objemu všetkých nových úverov domácnostiam.



Po znížení objemov novo poskytnutých úverov na bývanie v priebehu roku 2009 a ich medziročných poklesoch aj napriek znižovaniu úrokov, v roku 2010 začína opäť rásť záujem o úvery na bývanie.

Hoci sčasti to môže súvisieť aj s refinancovaním starších úverov, je to určitý náznak obnovovania aktivít na slovenskom trhu s bývaním.

Najbližší vývoj slovenského trhu s bývaním bude ovplyvňovaný predovšetkým dopytovými faktormi a naďalej bude regionálne značne diferencovaný. Výraznejšie pozitívne impulzy možno predpokladať najskôr v ekonomicky výkonnejších regiónoch (najmä Bratislavský kraj, Trnavský kraj a Žilinský kraj) a za predpokladu dosiahnutia trvalejšie udržateľného rastu budú ich postupne nasledovať aj ďalšie slovenské regióny.

4 Záver

Informačné potreby o slovenskom trhu s bývaním sú z pohľadu jeho aktérov značne diferencované a zatiaľ sú riešené skôr individuálnymi aktivitami zainteresovaných subjektov. Perspektívnym cieľom by malo byť postupné vytváranie efektívnejších databáz údajov o realitnom trhu pre široký okruh záujemcov.

Na základe vývoja viacerých vybraných ukazovateľov ohľadom trhu s bývaním na Slovensku možno zovšeobecniť, že aktuálne dochádza k jeho miernemu zotavovaniu a relatívnej stabilizácii. Ďalší vývoj slovenského trhu s bývaním bude závisieť od ekonomickej stability v Európe, čo je nevyhnutným predpokladom pre rast výkonnosti výrazne proexportne orientovanej slovenskej ekonomiky.

5 Literatúra:

[1] Cár, M.: Aktuálny stav a perspektívne možnosti zisťovania cien nehnuteľností na bývanie na Slovensku. In.: Slovenská štatistika a demografia 2/2010.

[2] Draft Technical Manual on Owner-Occupied Housing. Eurostat, February 2010.

Adresa autora:
Mikuláš Cár, Ing., PhD.
Národná banka Slovenska
Imricha Karvaša 1
813 25 Bratislava
Mikulas.Car@nbs.sk

Modelovanie cien rezidenčných nehnuteľností v USA
Modeling of residential housing prices in the US

Tomáš Cár

Abstract: The purpose of this article is to introduce a multilinear regression model for modeling of local housing prices at the metropolitan level in the US. The article starts with a discussion about the mortgage crisis that caused turbulences in the whole financial world. It continues with a description of a demand-supply equilibrium model that takes economic, demographic and housing variables into account when predicting housing prices at the MSA level. It also discusses specific techniques and methods, such as stepwise regression, time series, lagging independent variables, spline interpolation and extrapolation…etc., used in the model. Finally, this article presents results and their practical implications and concludes with suggestions for further research.

Key words: residential housing prices in the US, MSA, HPI, multi-linear econometric regression model, stepwise regression

Kľúčové slová: ceny rezidenčných nehnuteľností v USA, metropolitné štatistické územné jednotky, HPI, multilineárny ekonometrický regresný model, postupná regresia

JEL classification: R150 - Econometric and Input–Output Models; Other Models

1. Úvod

Realitný trh rezidenčných nehnuteľností v USA zažíval od deväťdesiatych rokov 20. storočia nebývalý rozmach aj vďaka (až príliš) dostupnému financovaniu. Banky rozdávali hypotéky „na počkanie“ aj pre menej bonitných klientov, ktorí by si ich za normálnych okolností nemohli dovoliť. Zdalo sa, že americký sen o tom, že každý môže bývať vo vlastnom dome stál na pevných základoch – veď dovtedy ceny rezidenčných nehnuteľností ešte nikdy nepoklesli na hodnote súčasne v celom USA, a hoci aj niektoré skutočnosti naznačovali možnú nestabilitu a hlavne neudržateľnosť stavu, nikto nevenoval alebo nechcel venovať pozornosť tikajúcej časovanej bombe. A tak zákonite prišiel tvrdý pád v podobe ekonomickej recesie, kedy sa na jednej strane trh presýtil novými nehnuteľnosťami a na strane druhej domácnosťami neschopnými splácať hypotéky, čo viedlo k rekordnému prepadu cien nehnuteľností a začarovanému kruhu neustále klesajúcich cien. Tento nevídaný prepad cien nehnuteľností nielenže drasticky zredukoval bohatstvo domácností v USA, kedy sa množstvo domácností ocitlo s väčším dlhom ako celkovými aktívami, ale takisto to katastroficky zasiahlo aj investorov po celom svete, ktorí investovali do amerických mortgage-backed securities. Je zrejmé, že nová, t.j. pokrízová, cena domu je jedným z kľúčových faktorov pre vlastníka hypotéky, na základe ktorých sa rozhodne, či ju bude môcť alebo vôbec chcieť ďalej splácať. Banky majú snahu pomôcť klientovi v splácaní hypotéky renegociovaním podmienok, ale to spravidla býva len odsúvaním problému do blízkej budúcnosti. V prípade, že klient nemôže alebo nechce splácať hypotéku, banke nezostáva nič iné ako speňažiť nehnuteľnosť, na ktorú bola hypotéka vydaná, a to je zväčša za podhodnotenú cenu, čo má za následok dve veci: Po prvé, ceny nehnuteľností sa dostávajú do začarovaného kruhu klesajúcich hodnôt, kedy sa neustále vytvára nová ponuka nehnuteľností a po druhé, ceny mortgage-backed securities strácajú na hodnote. Z uvedeného vyplýva akú dôležitú úlohu zohráva vývoj a predikcia cien nehnuteľností pre realitný trh i investičné stratégie. Nakoľko realitný trh v USA je špecifický pre každú geografickú oblasť, je na mieste analyzovať vývoj cien nehnuteľností jednotlivo pre každú oblasť na základe lokálnych

ekonomických fundamentov. Navyše dôsledky krízy sa signifikantne líšia v rôznych častiach USA napríklad aj vďaka rôznemu vystaveniu voči subprime hypotékam v daných regiónoch. USA je v súčasnosti rozdelené do 366 metropolitných štatistických jednotiek „metropolitan statistical areas“ (ďalej len MSAs). Tie reprezentujú základné urbanizačné jednotky, ku ktorým existujú popisné štatistické a ekonomické dáta. V nasledujúcom texte najprv všeobecne načrtnem model pre predikciu cien nehnuteľností, potom podrobnejšie predstavím jednotlivé vstupné dáta a štatistické metódy modelovania a na záver sa pokúsim zhodnotiť výsledky a praktický prínos prezentovaného modelu.

2. Použité dáta

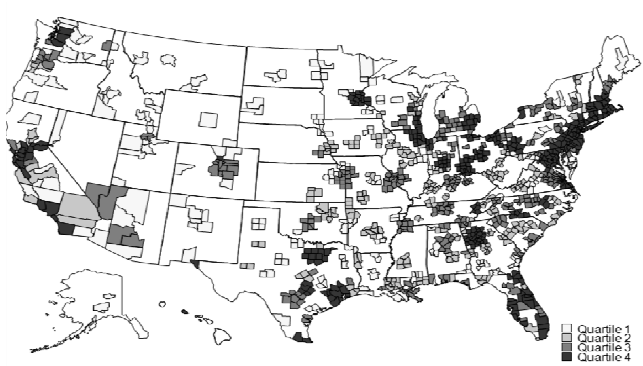
Na vytvorenie modelu, ktorý analyzuje a predpovedá vývoj budúcich cien rezidenčných nehnuteľností v USA bol použitý rovnovážny model dopytu a ponuky, ktorý berie do úvahy demografické a ekonomické premenné a parametre opisujúce realitný trh s rezidenčnými nehnuteľnosťami. Spomedzi množstva relevantných premenných som vybral také premenné, ktorých dostupnosť bola pre všetky MSAs a jednoznačne pomenúvali vzťah k dopytu alebo ponuke na realitnom trhu (Stevenson, 2008). Konkrétne povedané, použil som dáta zachytávajúce disponibilný príjem na obyvateľa, nezamestnanosť, hustotu obyvateľstva a vývoj migrácie ako základné premenné popisujúce dopytovú stránku. Na opis ponukovej stránky som si pomohol premennými zachytávajúcimi počet povolení na výstavbu nových domov (building permits), ktoré aj podľa Hwang a Quigley-ho (2006) reprezentujú veľmi dobrý odhad pre skutočný počet novo postavených domov (housing starts). Ďalej som použil dáta zachytávajúce prevládajúcu úrokovú mieru hypoték a cenový index rezidenčných nehnuteľností. Všetky premenné s výnimkou úrokových mier sa líšia pre rôzne MSAs. V nasledujúcej časti si detailnejšie priblížime použité dáta.

Cenový index rezidenčných nehnuteľností (Home Price Index, ďalej len HPI) som čerpal z FHFA (Fedeal Housing Finance Agency), ktorý štvrťročne zaznamenáva cenový index pre single-family rezidenčnú nehnuteľnosť na základe údajov z konformných hypoték od Freddie Mac a Fannie Mae. Použitý HPI index zachytáva všetky transakcie, teda aj nákup, aj refinancovanie rezidenčných nehnuteľností. Menšou nevýhodou je, že spomenutý HPI index využíva dáta iba z konformných hypoték, čiže chýbajú údaje o cenách rezidenčných nehnuteľností zakúpených napríklad subprime hypotékami. Tento nedostatok je však vyvážený hlavným benefitom FHFA HPI indexom, a to granularitou dát, čiže vieme rozlišovať cenovú úroveň rezidenčných nehnuteľností vo všetkých MSAs.

Počet povolení na výstavbu nových domov na úrovni MSA sú štvrťročné údaje zbierané od roku 1980 organizáciou US Census Bureau.

Dáta o disponibilnom príjme na obyvateľa pre rôzne MSAs v ročnej frekvencii publikuje úrad the Bureau of Economic Analysis (BEA).

Hustotu obyvateľstva (Obrázok 1) som získal podielom populácie a rozlohy každej MSA. Obe štatistiky ako aj údaje o migrácii obyvateľstva sú dostupné vďaka US Census Bureau. Ročné údaje o populácii sú dostupné na základe sčítavania obyvateľstva, ktoré sa vykonáva každých 10 rokov (posledné v roku 2010). V čase vykonávania tejto analýzy boli dostupné reálne údaje len z roku 2000, ale Census Bureau poskytuje aj projekcie populácií pre jednotlivé MSAs po roku 2000, ktoré som použil.



Obrázok 4: Hustota obyvateľstva v USA (údaje za December 2008)

Údaje o nezamestnanosti na úrovni MSA sú na mesačnej báze publikované úradom the Bureau of Labor Statistics (BLS).

Pre úrokovú mieru hypoték som sa rozhodol použiť jednotnú mieru pre všetky MSAs, a to úrokovú mieru pre typickú fixnú 30-ročnú hypotéku čerpanú zo zdroja HSH Financial Publishers. Tabuľka 1 porovnáva údaje nedávnych hodnôt nezávislých premenných pre 10 MSAs s najväčšími populáciami. Pre lepšiu predstavu o celkovom množstve vstupných dát si je treba uvedomiť, že hovoríme o databáze pozostávajúcej z 363 metropolitných jednotiek, z ktorých každá je charakterizovaná šiestimi ekonomickými alebo demografickými premennými o dĺžke vyše 80 pozorovaní v časovom rade (štvrťročné údaje od Q1 v roku 1990 do Q1 v roku 2010).

Tabuľka 2: Vstupné dáta pre 10 najľudnatejších MSAs za Q1 v roku 2010

MSA_NAME	Unempl	Permits	Income	Density	Migr
New York-Northern New Jersey-Long Isl.,NY-NJ-PA	9.53%	13,428	56,552	2,858	13,140
Los Angeles-Long Beach-Santa Ana, CA	11.90%	9,000	44,062	2,713	17,041
Chicago-Naperville-Joliet, IL-IN-WI	11.47%	5,040	46,693	1,354	925
Dallas-Fort Worth-Arlington, TX	8.47%	20,796	44,174	741	21,338
Houston-Sugar Land-Baytown, TX	8.60%	29,748	51,938	681	19,062
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	9.57%	6,724	50,057	1,269	-2,896
Atlanta-Sandy Springs-Marietta, GA	10.63%	8,700	36,583	674	16,744
Miami-Fort Lauderdale-Pompano Beach, FL	11.43%	4,996	44,540	1,073	-7,890
Washington-Arlington-Alexandria, DC-VA-MD-WV	6.83%	13,672	59,788	978	10,591
Boston-Cambridge-Quincy, MA-NH	8.83%	4,644	58,242	1,312	4,516

3. Špecifikácie modelu

Faktory, ktoré určujú budúci vývoj cien nehnuteľností v USA sú špecifické pre rôzne MSAs. V niektorých prípadoch môžu byť určujúce ekonomické fundamenty v jednotlivých oblastiach, v iných prípadoch môže ísť čisto o špekulatívne nákupy a očakávania, ktoré napríklad umelo zvýšia cenu. Z tohto dôvodu som sa rozhodol pridať do modelu aj nezávislú premennú HPI_{t-1}, ktorá reprezentuje oneskorenú premennú (lagged variable) závislej

premennej HPI. Týmto spôsobom testujem vysvetľovanie vývoja cien nehnuteľností ekonomickými fundamentami, ale aj cenovými očakávaniami účastníkov trhu.

Model predpovedajúci budúci vývoj cien rezidenčných nehnuteľností v USA predstavuje regresný model s viacerými nezávislými premennými a je založený na metóde dynamickej postupnej regresie (dynamic stepwise regression). Tá nám umožní automatický výber optimálnych, alebo len signifikantných premenných, kde model vďaka kombinovanému prístupu (backward and forward stepwise regression) zakaždým preveruje, či je možné vynechať niektorú z vysvetľujúcich premenných bez toho, aby sa podstatne zvýšil nevysvetlený súčet štvorcov odchýlok (residual sum of squares) v regresii.

V modeli používam dáta so štvrťročnou frekvenciou a keďže niektoré vysvetľujúce premenné sú k dispozícii len v podobe dát s ročnou frekvenciou, musel som využiť interpoláciu dát (uprednostnil som plastickejšiu cubic spline metódu pred lineárnou interpoláciou) na získanie štvrťročných dát. Podobne pre získanie najnovších dát, ktoré ešte neboli publikované (napríklad odhady populácie v rokoch 2009 alebo 2010), som využil techniku extrapolácie (spline function).

Navyše v regresnom modeli som použil aj časovú dummy variable (premennú), pomocou ktorej sa snažím zachytiť štrukturálne zmeny v modeli alebo zmenu vzťahov medzi premennými, ktoré mohli nastať v dôsledku ekonomickej krízy. Pred prasknutím cenovej bubliny na trhu nehnuteľností v USA mala HPI vo väčšine MSAs len stúpajúcu tendenciu. Po tomto období sa ceny prudko prepadli nadol. V modeli som preto použil túto dummy variable, a to nasledovným spôsobom: Od začiatku sledovaného obdobia, t.j. od Q1 1990 až po Q1 v roku 2006 som použil hodnotu 0 a pre hodnoty Q2 2006 a neskoršie som použil hodnotu 1.

V neposlednom rade som využil štandardnú techniku oneskorených nezávislých premenných, kedy som oneskoril všetky dopytové (nezamestnanosť, populácia a demografia) a jednu z ponukových nezávislých premenných (počet povolení na výstavbu nových domov) o dva štvrťroky. Tento krok je odôvodniteľný logikou, že momentálna zmena napríklad v nezamestnanosti alebo ponuke nových domov sa prejaví na cene nehnuteľností až o nejaký čas – v literatúre sa najčastejšie využíva oneskorenie o jeden, respektíve dva štvrťroky, ja som použil oneskorenie o dva štvrťroky. Vďaka uvedenej technike budem schopný okamžite predikovať vývoj cien nehnuteľností v nasledujúcich dvoch štvrťrokoch. V prípade, že by som vytvoril model na predikciu jednotlivých nezávislých premenných (napríklad pomocou metód časových radov, ARMA modelov), vedel by som predpovedať vývoj cien nehnuteľností do ešte hlbšej budúcnosti. Toto však predstavuje extra nadstavbu k súčasnému modelu, s ktorou sa v tomto príspevku zaoberať nebudem.

Zostavil som sústavu 363 rovníc, kde som skúmal vzťahy medzi nezávislými premennými a cenou nehnuteľností v jednotlivých MSAs. Výsledný všeobecný tvar modelu pre každú MSA vyzerá nasledovne:

$$HPI = \beta_0 + \beta_1 Income + \beta_2 Unemploy + \beta_3 PopDens + \beta_4 HousStart + \beta_5 30FRM + \\ + \beta_6 TIME + \beta_7 Migr + \beta_8 HPI_{lag} \quad , (1)$$

4. Výsledky modelu

Pri porovnaní výsledkov z 363 regresií mi vyšlo, že oneskorená premenná HPI_{lag} bola štatisticky významná až v 97% prípadoch. Nasledovaná bola ďalej časovou dummy premennou (69%), disponibilným príjmom na obyvateľa (68%), hustotou obyvateľstva (62%), počtom povolení na výstavbu nových domov (47%), nezamestnanosťou (41%) a úrokovou mierou hypoték (33%). Intuitívna interpretácia týchto výsledkov by mohla znieť, že ceny rezidenčných nehnuteľností boli predovšetkým poháňané cenovými očakávaniami ako aj to, že vzťahy ktoré vysvetľovali vývoj cien pred rokom 2006 a po ňom boli odlišné.

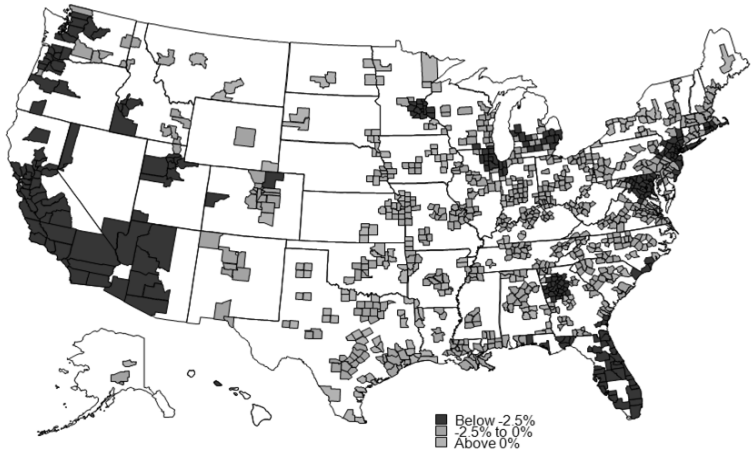
Spomedzi ekonomických fundamentov, disponibilný príjem na obyvateľa a hustota obyvateľstva sa zdajú byť najlepšími premennými, ktoré vysvetľujú pohyb a vývoj cien nehnuteľností.

Typickým výsledkom analýzy je predikcia vývoja budúceho vývoja cien nehnuteľností (Tabuľka 1 a Obrázok 2). Uvedené predikcie môžu pomôcť pri rozhodovaní sa tak o otázkach o kúpe alebo predaji nehnuteľností, ako aj pri oceňovaní mortgage-backed securities, kedy očakávaná cena nehnuteľností výrazne ovplyvňuje pravdepodobnosť defaultu hypotéky, čo má priamy dopad na cenu ceniny/dlhopisu.

Tabuľka 2: Predpoveď vývoja HPI na Q2 a Q3 v roku 2010 pre 10 najľudnatejších MSAs

MSA	HPI	HPA (%)		
	10Q1	10Q1	10Q2f	10Q3f
New York-Northern New Jersey-Long Isl.,NY-NJ-PA	234.04	-0.94	-2.59	-2.66
Los Angeles-Long Beach-Santa Ana, CA	240.1	-0.85	-1.61	-1.22
Chicago-Naperville-Joliet, IL-IN-WI	159	-1.91	-2.65	-2.66
Dallas-Fort Worth-Arlington, TX	163.23	-0.61	-1.21	-0.98
Houston-Sugar Land-Baytown, TX	189.78	-0.86	-1.21	-1.28
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	198.42	-0.94	-1.36	-1.42
Atlanta-Sandy Springs-Marietta, GA	163.19	-1.71	-2.95	-2.86
Miami-Fort Lauderdale-Pompano Beach, FL	197.22	-0.68	-3.78	-2.98
Washington-Arlington-Alexandria, DC-VA-MD-WV	217.66	-0.93	-1.95	-1.57
Boston-Cambridge-Quincy, MA-NH	223.68	-0.78	-0.48	-0.47

Z Obrázka 2, kde sú znázornené predpovede vývoja cien rezidenčných nehnuteľností, je jasne vidieť odlišnosť jednolivých geografických oblastí. V tomto prípade môžeme vidieť, že najmä oblasti štátov Kalifornie a Floridy mali najhoršie výhliadky cenového vývoja. Každý švrťrok sa model rekalibruje a takisto aj vyhodnocuje presnosť jeho predpovedí pomocou viacerých metód – absolútna a relatívna odchýlka, správnosť smeru predikcie…atď.



Obrázok 2: Predpoveď vývoja cien nehnuteľností na Q2 v roku 2010

5. Záver

Cieľom tohto článku bolo prezentovať multilineárny regresný model na predpovedanie vývoja cien rezidenčných nehnuteľností na metropolitných úrovniach v USA. Problematika cien nehnuteľností v USA je výnimočná svojou komplexnosťou, preto domnievať sa, že na svete existuje nejaký model, ktorý by bol schopný perfektne vysvetliť pohyby cien nehnuteľností v USA je prinajmenšom naivné. Napriek tomu si myslím, že prezentovaný model ponúka solídne riešenia v podobe predpovedí vývoja cien rezidenčných nehnuteľností v metropolitných oblastiach USA v blízkej budúcnosti, a to tak, že na jednej strane ponúka vysvetlenie pohybu cien lokálnymi ekonomickými a demografickými fundamentami a na strane druhej ponúka aj vysvetlenie pomocou očakávaní budúcich cien. Z prezentovaných výsledkov vyplýva, že očakávania budúcich cien ovplyvňujú vývoj budúcich cien vo veľkej miere a spomedzi ekonomických a demografických fundamentov disponibilný príjem na obyvateľa a hustota obyvateľstva sa javia ako štatisticky najvýznamnejšie premenné. Presnosť modelu je ľahko overiteľná každý nový štvrťrok a pravidelná rekalibrácia parametrov by mala zabezpečiť dynamickú adaptáciu modelu na aktuálne podmienky. Ako som už načrtol v časti špecifikácie modelu, model môže byť obohatený o predikcie nezávislých premenných, čím by sa v konečnom dôsledku zvýšil počet možných predikcií vývoja cien nehnuteľností do budúcnosti.

6. Literatúra

[1] STEVENSON, S. 2008. Modeling housing market fundamentals: Empirical evidence of extreme market conditions. In: Real Estate Economics, Vol. 36, 2008, pp. 1 – 29.

[2] EPPLI, M. J., SHILLING, J.D., VANDELL, K.D. 1998. What moves retail property returns at the metropolitan level. In: Journal of Real Estate Finance and Economics, Vol. 16:3, 1998, pp. 317 – 342.

[3] HWANG, M., QUIGLEY, J.M. 2006. Economic fundamentals in local housing markets: Evidence from U.S. metropolitan regions. In: Journal of regional science, Vol. 46, No. 3, 2006, pp. 425 – 453.

Adresa autora:

Tomáš Cár, Mgr., MSc., MBA
Alžbetin Dvor 631
900 42 Miloslavov
tomascar@yahoo.com

Passing-Bablokova regresní metoda a návrh intervalových odhadů regresních koeficientů

Passing-Bablok's Regression Metod and Propsal of Interval Coeficient Estimates

Jindřich Černý

Abstract: The article focuses on comparison of robust regression analysis methods and least squares method namely in models with one explanatory variable and one dependent variable. Especial focus is given to Passing- Bablok regression and problems with implementing to software. Robust methods offer the possibility of both high-quality estimation and at the same time robust estimates. To demonstrate fitness and juxtaposition, simulated data similar to experimental one was used.

Key words: Passing-Bablok regression, outlier, leverage point, robust regression, break point.

Klíčová slova: Passink-Bablok regrese, odlehlá pozorování, vlivné body, robustní regrese, bod zvratu.

JEL classification: C2, C3

1. Úvod

V mnoha úlohách regresní analýzy je zapotřebí použít pro odhad parametrů jinou metodu než běžně používanou metodu nejmenších čtverců (MNČ), která je téměř všude používána, pro mnohé praktické použití ale nejsou odhady provedené touto metodou dostačující. Zejména předpoklady o normálním rozdělení chyb většinou nebývají dodrženy, vybočující pozorování ať již leverages (zásadně odlišné hodnoty ve vysvětlujících proměnných) či outliers (zásadně odlišné hodnoty ve vysvětlované proměnné) nebývají výjimkou, taktéž chybová rozdělení s „těžšími" konci nejsou u reálných dat nic neobvyklého. Nehodící se měření nebo body nelze jednoduše vyškrtnout a stává se pak nutností použít některou metodu odolnou proti vybočujícím pozorováním tedy metodu robustní.

Jednou z takových metod je Passing-Bablokova regrese popsaná v [3], poprvé byla zmíněna v [4]. Tato metoda je hojně používaná v medicínském a farmaceutickém softwaru při měření různých výstupních obsahů látek v závislosti na vstupních dávkách. Nejčastěji je používána při srovnávání dvou měřících metod – součástí tohoto softwaru jsou pak testy na směrnici rovnu jedné a absolutní člen rovný nule (pro srovnávání metod), proto je zřejmě vždy kladen důraz hlavně na použití u přímky. V nalezené literatuře není bližší popis algoritmu. Cílem tohoto článku je představit tuto metodu, prokázat její robustnost, odhadnout bod zvratu a na příkladě ukázat její využití.

2. Passing-Bablokova regrese - popis

Princip Passing-Bablokovy regrese je velmi jednoduchý. Je popsaný v [3] a stručně v [2]. Nejprve uvažujme model přímky $y = b_0 + b_1 x$, z naměřených hodnoty x a y vytvoříme všechny kombinace dvou bodů, tedy dvojice x_i, y_i a x_j, y_j, tyto dvojice bodů spojíme přímkou a každé přímce určíme její směrnici, takže po úpravě pro každou směrnici dostáváme

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j} = 0 \text{ pro } 1 \le i < j \le n, \tag{1}$$

kde n je počet pozorování. Dále pak určíme medián ze všech takto vypočtených hodnot b_{ij} a ten použijeme jako odhad b_1 tedy

$$\hat{b}_1 = \tilde{b}_{ij}, \tag{2}$$

Pro odhad b_0 použijeme medián rozdílů

$$\hat{b}_0 = a_i \text{ kde } a_i = y_i - \hat{b}_1 x_i. \tag{3}$$

Většinou potřebujeme také intervalový odhad parametrů. Jedna z možností je použít některou z metod pro intervalový odhad mediánu, takže horní a dolní hranice pro b_1 je

$$R_L = \frac{z - u_{\alpha/2}\sqrt{z}}{2} \text{ a } R_H = \frac{z + u_{\alpha/2}\sqrt{z}}{2} + 1, \tag{4}$$

Kde R_L a R_H je pořadí ze vzestupně seřazených hodnot odhadnutých směrnic, u je kvantil normálního rozdělení a z je celkový počet odhadnutých směrnic ze všech dvojic bodů, je to kombinace

$$z = \binom{n}{2}. \tag{5}$$

Pro b_o je intervalový odhad

$$R_L = \frac{n - u_{\alpha/2}\sqrt{n}}{2} \text{ a } R_H = \frac{n + u_{\alpha/2}\sqrt{n}}{2} + 1, \tag{6}$$

kde R_L a R_H je pořadí ze vzestupně seřazených hodnot odhadnutých b_0, u je kvantil normálního rozdělení a n je počet pozorování. Tyto vzorce 4 a 6 můžeme použít pro n>20.

Tato robustní metoda je neparametrická a řadí se do metod s vyšším bodem zvratu, je odolná proti leverages i outliers.

3. Problémy při výpočtu

Celkový počet odhadnutých směrnic je kombinace viz vzorec 5. Takže prvním problémem je vysoký počet odhadnutých směrnic, jak ukazuje tabulka 1.

Tabulka 3: Celkový počet odhadnutých směrnic

n	5	10	50	100	500	1000
Počet směrnic	10	45	1225	4950	124750	499500

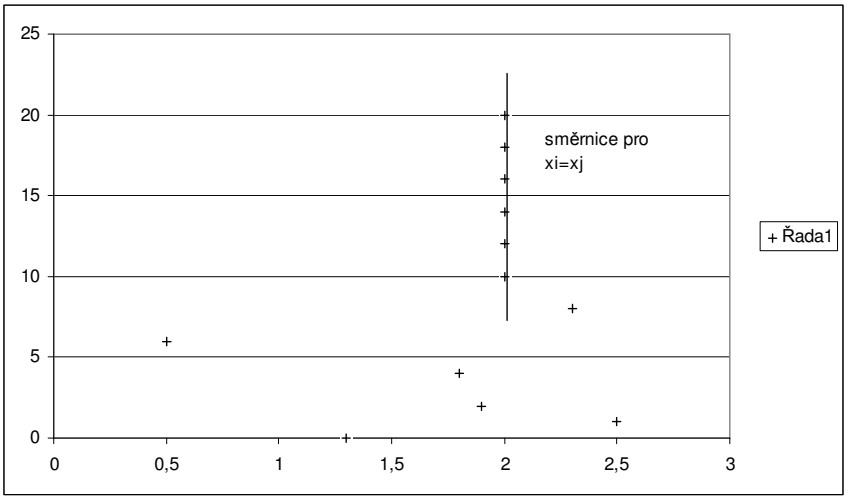
Ve vícerozměrných úlohách by byl tento počet ještě vyšší. Dnešní výpočetní technika takovou velikost úloh zvládá, horší je použití v jednočipových aplikacích nebo v řídících jednotkách apod., kde výpočetní síla nemusí být dostatečná.

Další významný problém nastává v okamžiku, kdy $x_i = x_j$ (ve vzorci to znamená dělení nulou). V reálných datech se tento problém může vyskytnout – pro stejnou hodnotu vysvětlující proměnné naměříme různé hodnoty výstupu. V literatuře není uveden žádný popis, jak v takovém případě postupovat.

K řešení tohoto problému jsou možné dva přístupy. První možnost je odstranění „špatné“ dvojice bodů. Druhou možností je nahrazení směrnice jinou hodnotou. Pro druhou možnost můžeme teoreticky předpokládat, že směrnice se blíží k $-\infty$ nebo $+\infty$, jak je ukázáno na obrázku 1.

Návrh na řešení tohoto problému je jednoduchý – jestliže většina směrnic je větší než 0, nahradíme „špatné“ dvojice největší vypočítanou směrnicí. Naopak je-li většina odhadnutých

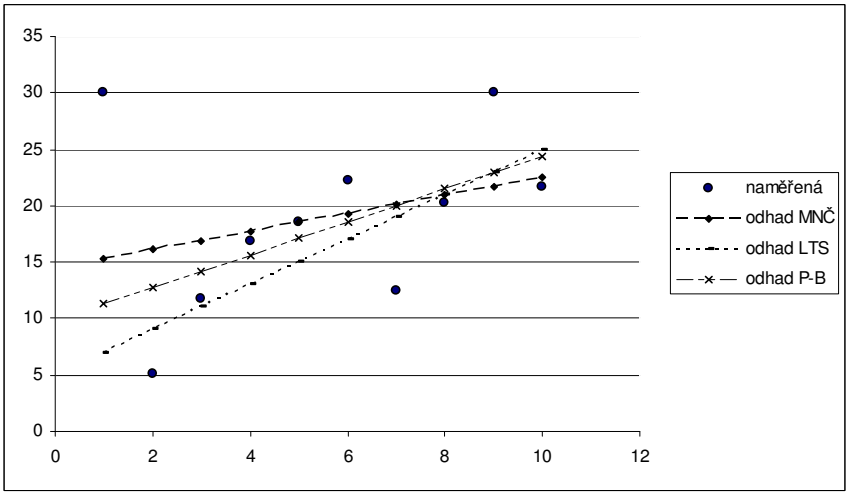
směrnic menších, nahradíme tyto dvojice nejmenší vypočtenou hodnotou. Tím v podstatě posouváme medián nahoru nebo dolů.



Obrázek 5: Data kde se vyskytují body $x_i = x_j$

4. Příklad odhadů

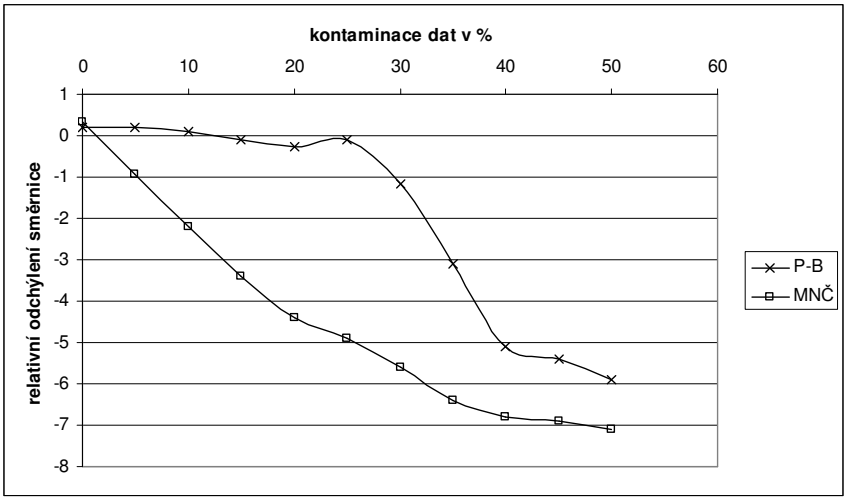
Odhady byly provedeny v MS excelu, za použití jednoduché procedury popsané v [2]. Pro ilustraci bylo provedeno srovnání odhadů Passing-Bablokovi regrese, MNČ a metody nejmenších useknutých čtverců (LTS) viz obrázek 2.



Obrázek 6: Odhady různými metodami

Rozdíl mezi Passing-Bablokovou metodou a LTS je způsoben malým počtem pozorování, při větších výběrech dávají tyto metody podobné výsledky.

Na obrázku 3 je vidět selhání metody při zvětšující se kontaminaci dat, pro ilustraci byla použita data s kontaminací náhodné složky Cauchyho rozdělením viz [1], model je postupně kontaminován větší chybou ve vysvětlující proměnné, až dojde k selhání odhadu směrnice, to je vyjádřeno relativní chybou. Na MNČ je vidět, že k selhání dojde při první kontaminaci modelu, odlehlé pozorování má tak velký vliv, že vychýlí odhady směrnice. Na průběhu selhání Passing-Bablokovi regrese je vidět, že v tomto případě k němu dochází až při poměrně vysoké kontaminaci modelu (okolo25%). Výsledky odhadů je možno použít i k identifikaci odlehlých pozorování, tak jako u jiných metod s vysokým bodem zvratu. Odhady absolutního členu vykazují menší stabilitu než odhady směrnice. Zvláště při použití intervalových odhadů.



Obrázek 7: Odhady různými metodami

5. Závěr

V příspěvku bylo ukázáno, že popsaná Passing-Bablokova regrese je metodou robustní. Simulacemi bylo zjištěno, že tato regresní metoda je odolná proti outliers i leverages. Problémem jsou stejné hodnoty ve vysvětlované proměnné, v případě neošetření algoritmu dochází k dělení nulou a chybovým hláškám. Experimentálně byl odhadnut bod zvratu mezi 25% až 35%. Tato metoda je velmi jednoduchá a dává velmi dobré výsledky. Při větším výběru jsou pak i intervalové odhady v rozumných mezích. Po provedení velkého počtu simulací na různě kontaminovaných datech a různě velkých výběrech, lze konstatovat, že odhady jsou velmi stabilní. Tato metody by se měla zcela jistě zařadit, mezi běžně nabízené metody ve statistickém software.

Dále bylo vyzkoušeno, že odhady z této metody mohou být velmi dobře použity jako výchozí odhady pro jiné iterační metody, například MM-odhady.

Další modifikace této metody mohou být například odhady regresních parametrů pro kvadratický či jiný model.

Cílem další práce bude posoudit, zda některá kritéria pro výběr robustních regresních modelů například AICC a BICC jsou vhodná pro použití s touto metodou a případně odvodit a použít další diagnostické prvky pro tuto metodu, tak aby se dala plnohodnotně použít.

6. Literatura

[1]ČERNÝ, J. 2005. Srovnání vybraných metod robustní regrese s použitím SAS 9.1. Praha: VŠE 2005

[2]ČERNÝ, J. 2010. Vybrané aspekty robustní regrese. Sborník dne doktorandů FIS. Praha: VŠE 2010

[3]DOHNAL, L. 2070. Klasická nebo robustní lineární regrese? Praha: ÚKBLD VFN a 1.LF UK 2007

[4]PASSING, H. – BABLOK, W. et al. 1988. A General Regression Procedure for Method Transformation. Application of Linear Regression Procedures for Method Comparison Studies in Clinical Chemistry, Part III, Clinical Chemistry and Laboratory Medicine, 26, no. 11 p. 783-790, Mannheim: Walter de Gruyter GmbH & Co., 1988

Adresa autora:

Jindřich Černý, Ing.
VŠE Praha
nám. W.Churchilla 4, 130 67 Praha 3
Česká republika
cernyj@vse.cz

Poděkování: Tento příspěvek vznikl za podpory grantu IGA 26/2010.

Lidské zdroje v ČR současnost a perspektivy
Human Resources in the Czech Republic – Present Time and Prospects

Tomáš Fiala, Jitka Langhamrová

Abstract: The paper describes the supposed development of the population of the Czech Republic in productive age in the next fifty years. The population projection is based on latest available data. The productive age is considered since 20 until the retirement age which is increasing in time. Not only the development of the number but also the development of the proportion and age structure of the population of productive age is analyzed. The projection indicates decrease of the number and proportion of people in productive age as well as population ageing of the labor force.

Key words: productive age, population projection, population ageing.

Klíčová slova: produktivní věk, populační projekce, stárnutí populace.

JEL classification: J11, J21, J24

1. Úvod

Pro ekonomický vývoj každé země je důležitý relativní dostatek nejen například finančních či materiálních zdrojů, ale i zdrojů lidských, zdrojů pracovních sil, tj. obyvatelstva v produktivním věku. Za dolní hranici produktivního věku je považován obvyklý věk zahájení ekonomické aktivity, což je v dnešní době přibližně 20 let. Jeho horní hranicí pak je obvyklý věk odchodu do důchodu. Protože v současné době v České republice dochází k prodlužování důchodového věku, není tato hranice v čase konstantní, ale rostoucí, odpovídající současné právní úpravě.

Na rozdíl od vývoje úmrtnosti, který je poměrně stabilní (střední délka života trvale roste) dochází v případě plodnosti a migrace v ČR k relativně častým a poměrně výrazným změnám trendů vývoje. Málokdo asi očekával tak prudký pokles plodnosti ve druhé polovině devadesátých let a rovněž její opětovný nárůst počátkem tohoto tisíciletí byl pro mnohé až překvapivě vysoký. V roce 2009 však byla plodnost o něco nižší než v roce 2008 a rovněž předběžná data za první polovinu roku 2010 naznačují pokračující stagnaci plodnosti.

Ještě větší změny vykazuje vývoj migrace. Zatímco během devadesátých let se pohyboval migrační přírůstek kolem 10 tisíc osob ročně, v letech 2005–2006 dosahoval hodnot již kolem 35 tisíc a v letech 2007–2008 dokonce 70–80 tisíc osob ročně. V roce 2009 však pokles na méně než 30 tisíc a v prvním pololetí roku 2010 činil pouze o něco málo více než 3 tisíce osob.

Vývoj plodnosti a migrace pochopitelně výrazným způsobem ovlivňuje vývoj demografické struktury obyvatelstva, tedy i vývoj počtu, podílu a demografické struktury obyvatelstva v produktivním věku. Ve svém příspěvku analyzujeme současný stav populace v produktivním věku v ČR a její předpokládaný vývoj během následujících 50 let.

2. Metodologické poznámky

Prahovou demografickou strukturou pro výpočet projekce bylo aktuální demografické složení obyvatelstva ČR podle věku a pohlaví k 31. prosinci 2009 (tj. k 1. lednu 2010) publikované ČSÚ ([1]). Předpokládaný vývoj do roku 2060 byl vypočten na základě pracovní verze aktuální populační projekce ČR vypracované na katedře demografie FIS VŠE v roce 2010.

Vzhledem ke stagnaci plodnosti v posledním období vycházela projekce z předpokladu, že úhrnná plodnost českých žen poroste v tomto desetiletí mnohem pomaleji než v předchozím období a do roku 2020 dosáhne hodnoty pouze 1,55. V dalším období se předpokládal lineární růst úhrnné plodnosti k hodnotě 1,70 v roce 2060. O struktuře plodnosti se předpokládalo, že se do roku 2020 přiblíží struktuře plodnosti nizozemských žen, kde již byla ukončena transformace plodnosti do vyššího věku a plodnost je zde poměrně stabilní.

O úmrtnosti se předpokládalo, že střední délka života mužů i žen nadále poroste po celé prognózované období s tím, že roční nárůst se bude postupně snižovat ze současných hodnot 0,30 pro muže, resp. 0,25 pro ženy na 0,20 pro muže resp. 0,17 pro ženy na konci prognózovaného období. V roce 2060 se předpokládá střední délka života mužů 85,5 roku, střední délka života žen 89,5 roku. Předpokládalo se zachování současné struktury úmrtnosti, tj. konstantní relativní pokles pravděpodobností úmrtí pro všechny jednotky věku.

Vzhledem k nízké migraci v posledním období byl přijat předpoklad, že migrační přírůstek bude činit pouze 10 tisíc osob ročně. Současně se předpokládal nárůst podílu žen a dětí mezi imigranty.

3. Přehled hlavních výsledků

Vývoj počtu a věkové struktury celé populace

Podle předpokládaného scénáře dalšího demografického vývoje by počet obyvatel České republiky ještě zhruba 10 let rostl, kolem roku 2020 by překročil 10 600 000 osob. V dalších letech by se však projevily důsledky poklesu počtu žen v reprodukčním věku (a tedy i poklesu počtu živě narozených), poměrně nízké plodnosti a nepříliš vysokého migračního přírůstku. Počet obyvatel ČR by začal trvale klesat, přibližně v roce 2030 by se vrátil na současnou úroveň. Pokles by však dále pokračoval, koncem čtyřicátých let by počet obyvatel ČR pokles pod 10 milionů a v roce 2060 by byl pravděpodobně jen o málo vyšší než 9,5 milionu s perspektivou dalšího poklesu.

Tabulka 4: Vývoj počtu obyvatel a velikosti ekonomických generací ČR

Rok	Počty osob				Podíly (v % z celkového počtu obyvatel)			
	předpro-duktivní věk	produktivní věk	popro-duktivní věk	celkem	předpro-duktivní věk	produktivní věk	popro-duktivní věk	celkem
2010	2 110 361	6 100 731	2 295 721	10 506 813	20,1	58,1	21,8	100,0
2015	2 065 608	6 089 906	2 430 908	10 586 421	19,5	57,5	23,0	100,0
2020	2 130 270	5 963 381	2 522 457	10 616 109	20,1	56,2	23,8	100,0
2025	2 129 922	5 926 165	2 529 101	10 585 187	20,1	56,0	23,9	100,0
2030	1 981 921	5 945 484	2 555 756	10 483 161	18,9	56,7	24,4	100,0
2035	1 818 770	5 885 929	2 636 462	10 341 160	17,6	56,9	25,5	100,0
2040	1 730 689	5 587 563	2 881 357	10 199 609	17,0	54,8	28,2	100,0
2045	1 712 885	5 200 947	3 151 810	10 065 642	17,0	51,7	31,3	100,0
2050	1 732 619	4 904 406	3 282 449	9 919 474	17,5	49,4	33,1	100,0
2055	1 732 838	4 643 533	3 371 232	9 747 602	17,8	47,6	34,6	100,0
2060	1 677 309	4 497 833	3 371 541	9 546 683	17,6	47,1	35,3	100,0

Zdroj: Vlastní projekce na základě výchozích dat ČSÚ

Klesal by rovněž počet osob v produktivním věku, a to navzdory předpokládanému postupnému zvyšování důchodového věku na 65 let do roku 2030. Zatímco v současné době činí počet osob v produktivním věku zhruba 6 100 000 osob, za 10 let by byl již o něco nižší než 6 milionů, na začátku 40. let by poklesl po 5,5 milionu a v roce 2060 by počet osob v produktivním věku byl již nižší než 5 milionů. Klesal by rovněž podíl osob v produktivním věku z celkového počtu obyvatel. Ze současných více než 58 % by se do konce 30. let snížil na 55 %. V následujících deseti letech, kdy do důchodu postupně odejdou početné skupiny osob narozené v 70. letech, lze očekávat pokles podílu osob v produktivním věku pod 50 % celkového počtu obyvatel. (Viz Tab. 1.). Naproti tomu dojde ke zvýšení podílu osob

v poproduktivním věku ze současných 21,8 % na více než 35 %. Stárnutí populace ČR bude pokračovat.

Vývoj počtu a věkové struktury osob v produktivním věku

Stárnutí obyvatelstva se nevyhne ani populaci v produktivním věku. Kromě snižování počtu i podílu osob v produktivním věku se bude měnit i věková struktura těchto osob. Bude se snižovat podíl mladších osob a naopak zvyšovat podíl osob ve vyšších věkových kategoriích produktivního věku. Hlavní příčinou je především velmi nerovnoměrná věková struktura populace ČR; vysoký podíl osob narozených v 70. letech a naopak nízký podíl osob narozených na přelomu tisíciletí. V současné době jsou nejpočetnější pětiletou věkovou skupinou 30–34letí, jejich počet je vyšší než 900 tisíc; přitom počet osob o 20 let mladších (10–14letí) je zhruba poloviční.

Počet osob v mladším produktivním věku (20–39letí), který je v současné době vyšší než 3 200 000 se během následujících 15 let sníží téměř o milion a po roce 2050 může být počet osob v této věkové kategorii nižší než 2 miliony. Naproti tomu počet osob v produktivním věku 40letých a starších do roku 2035 téměř o milion vzroste, pak však i v této věkové skupině nastane pokles počtu osob. Podrobnější údaje o vývoji počtu osob v jednotlivých pětiletých skupinách produktivního věku udává Tab. 2.

Tabulka 2: Vývoj věkového složení osob v produktivním věku v ČR (počty osob)

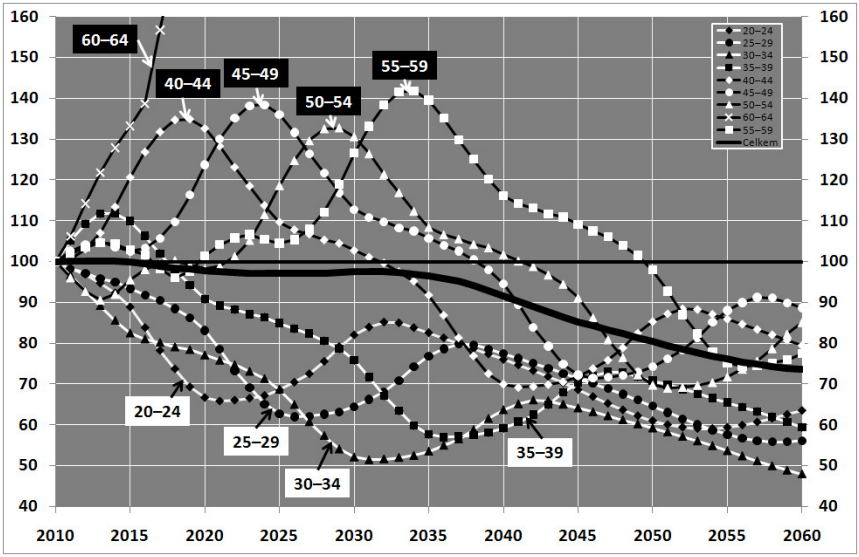
Rok	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	Celkem
2010	700 740	758 921	927 504	845 964	700 861	679 938	700 751	643 490	142 562	6 100 731
2015	622 308	708 665	764 668	929 951	845 367	697 247	669 909	661 864	189 927	6 089 906
2020	467 186	630 665	714 779	768 024	929 298	841 263	688 082	652 913	271 171	5 963 381
2025	480 256	476 070	637 152	718 544	768 706	925 166	831 354	672 324	416 593	5 926 165
2030	574 503	489 275	483 032	641 334	719 828	766 604	915 224	814 424	541 261	5 945 484
2035	578 404	583 514	496 316	487 753	643 204	718 635	759 746	898 092	720 265	5 885 929
2040	532 467	587 549	590 487	501 082	490 405	642 903	713 131	746 987	782 553	5 587 563
2045	480 666	541 811	594 600	595 122	503 810	491 350	638 838	702 381	652 368	5 200 947
2050	427 428	490 199	549 012	599 286	597 664	504 937	489 291	630 239	616 350	4 904 406
2055	416 068	437 135	497 538	553 842	601 908	598 578	503 194	483 686	551 585	4 643 533
2060	444 975	425 888	444 598	502 498	556 685	603 008	596 520	498 046	425 615	4 497 833

Zdroj: Vlastní projekce na základě výchozích dat ČSÚ

Změny počtu osob v jednotlivých věkových skupinách produktivního věku můžeme vyjádřit pomocí indexů porovnávajících vývoj počtu osob vzhledem k výchozím počtům na počátku roku 2010. Zatímco počet osob ve věkových skupinách produktivního věku do 35 let bude již v tomto desetiletí klesat, u vyšších věkových kategorií zaznamenáme nejprve růst (jak do tohoto věku budou postupně vstupovat početné skupiny osob narozené v 70. letech), pokles se dostaví až později. Jedinou výjimkou je pochopitelně počet 60letých a starších osob, vzhledem k tomu, že důchodový věk se postupně zvyšuje nad 60 let, bude se zvyšovat i počet osob produktivního věku 60letých a starších. (Viz Obr. 1.).

Bude se tedy měnit nejen počet, ale i věkové složení osob produktivního věku. Zatímco v současné době je více než polovina osob produktivního věku mladších 40 let, během 20 let poklesne podíl osob v této věkové kategorii téměř na třetinu počtu všech osob produktivního věku, v dalších letech se zvýší zhruba na 40 %. Podíl osob produktivního věku starších 40 let naopak poroste. Vzhledem ke zvyšování důchodového věku zaznamenáme nejvyšší růst u podílu osob produktivního věku starších 60 let. Zatímco v současné době jsou pouze 2,3 % osob produktivního věku starší 60 let, kolem roku 2040 se může pohybovat podíl těchto osob na hodnotách blízkých 14 %. Podrobnější údaje o změnách věkové struktury osob produktivního věku zachycuje Tab. 3. a Tab. 4. a Obr. 2.

O stárnutí populace v produktivním věku svědčí i růst průměrného věku osoby v produktivním věku. Zatímco v současné době je jeho hodnota o něco nižší než 40, v nejbližších letech tuto hranici překročí a ve 30. letech se může průměrný věk osob v produktivním věku pohybovat okolo 44 let. (Viz Obr. 3)



Zdroj dat: Vlastní projekce na základě výchozích dat ČSÚ

Obr. 1: Vývoj indexů počtu osob v produktivním věku

Tabulka 3: Vývoj věkového složení osob v produktivním věku v ČR (podíly v % celkového počtu osob v produktivním věku)

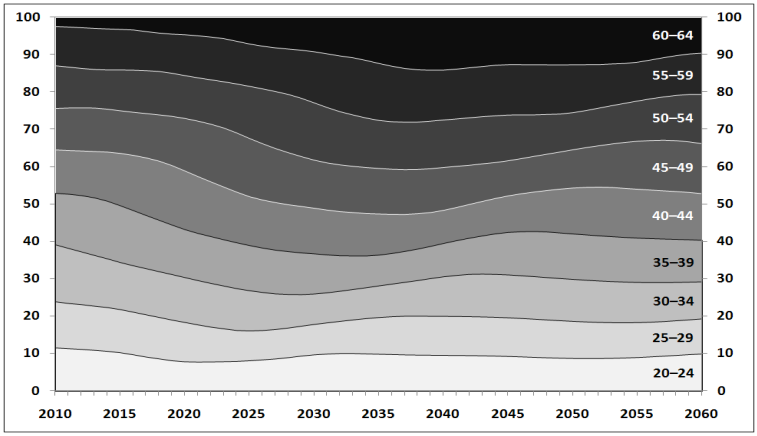
Rok	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64	Celkem
2010	11,5	12,4	15,2	13,9	11,5	11,1	11,5	10,5	2,3	100,0
2015	10,2	11,6	12,6	15,3	13,9	11,4	11,0	10,9	3,1	100,0
2020	7,8	10,6	12,0	12,9	15,6	14,1	11,5	10,9	4,5	100,0
2025	8,1	8,0	10,8	12,1	13,0	15,6	14,0	11,3	7,0	100,0
2030	9,7	8,2	8,1	10,8	12,1	12,9	15,4	13,7	9,1	100,0
2035	9,8	9,9	8,4	8,3	10,9	12,2	12,9	15,3	12,2	100,0
2040	9,5	10,5	10,6	9,0	8,8	11,5	12,8	13,4	14,0	100,0
2045	9,2	10,4	11,4	11,4	9,7	9,4	12,3	13,5	12,5	100,0
2050	8,7	10,0	11,2	12,2	12,2	10,3	10,0	12,9	12,6	100,0
2055	9,0	9,4	10,7	11,9	13,0	12,9	10,8	10,4	11,9	100,0
2060	9,9	9,5	9,9	11,2	12,4	13,4	13,3	11,1	9,5	100,0

Zdroj: Vlastní projekce na základě výchozích dat ČSÚ

Tabulka 4: Vývoj věkového složení osob v produktivním věku v ČR (kumulované podíly v % celkového počtu osob v produktivním věku)

Rok	do 25 let	do 30 let	do 35 let	do 40 let	do 45 let	do 50 let	do 55 let	do 60 let	Celkem
2010	11,5	23,9	39,1	53,0	64,5	75,6	87,1	97,7	100,0
2015	10,2	21,9	34,4	49,7	63,6	75,0	86,0	96,9	100,0
2020	7,8	18,4	30,4	43,3	58,9	73,0	84,5	95,5	100,0
2025	8,1	16,1	26,9	39,0	52,0	67,6	81,6	93,0	100,0
2030	9,7	17,9	26,0	36,8	48,9	61,8	77,2	90,9	100,0
2035	9,8	19,7	28,2	36,5	47,4	59,6	72,5	87,8	100,0
2040	9,5	20,0	30,6	39,6	48,4	59,9	72,6	86,0	100,0
2045	9,2	19,7	31,1	42,5	52,2	61,7	74,0	87,5	100,0
2050	8,7	18,7	29,9	42,1	54,3	64,6	74,6	87,4	100,0
2055	9,0	18,4	29,1	41,0	54,0	66,9	77,7	88,1	100,0
2060	9,9	19,4	29,2	40,4	52,8	66,2	79,5	90,5	100,0

Zdroj: Vlastní projekce na základě výchozích dat ČSÚ



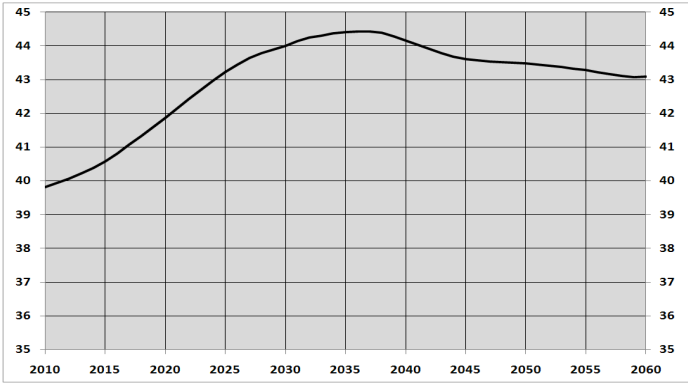
Zdroj dat: Vlastní projekce na základě výchozích dat ČSÚ

Obr. 2: Vývoj věkové struktury osob v produktivním věku

4. Závěry

Prezentované výsledky platí za předpokladů, na jejichž základě byla projekce vypočtena. Řada předpokladů je velmi zjednodušených. Není například vyloučena možnost dalšího zvyšování důchodového věku nad hranici 65 let, i když je otázkou, zda by pro lidi tohoto věku byl dostatek vhodných pracovních příležitostí. Rovněž je možné, že dojde k oživení a nárůstu migračního salda nad předpokládaných 10 tisíc osob ročně či k rychlejšímu nárůstu plodnosti.

V každém případě je však třeba se nad eventuálními důsledky souvisejícími se snižováním podílu a stárnutím populace produktivního věku zamýšlet co nejdříve a hledat jejich řešení s dostatečným předstihem. Tím se zvýší naděje, že chystaná opatření, která nemusejí být pochopitelně vždy populární a vítaná, vzniknou na základě serióznía podrobné analýzy, celospolečenské diskuse a s ohledem na zájmy všech společenských vrstev. Sníží se tak riziko že by tato opatření vyvolala další snížení životní úrovně osob s nižšími příjmy, měla za následek růst sociálních rozdílů a zvýšila tak napětí ve společnosti.



Zdroj dat: Vlastní projekce na základě výchozích dat ČSÚ

Obr. 3: Vývoj průměrného věku osob v produktivním věku

5. Literatura

[1]BOGUE, D. J., ARRIAGA, E. E., ANDERTON, D., L. (eds.). 1993. Readings in Population Research Methodology Vol. 5. Population Models, Projections and Estimates. United Nations Population Fund, Social Development Center, Chicago, Illinois.

[2]FIALA, T. 2006 Dva přístupy modelování vývoje úmrtnosti v populační projekci a jejich aplikace na populaci ČR. Bratislava 05.10.2006 – 06.10.2006. In: Forum Statisticum Slovacum 4/2006. Bratislava : Slovenská štatistická a demografická spoločnosť, s. 44–55. ISSN 1336-7420.

[3]FIALA, T., LANGHAMROVÁ, J. 2009. Human resources in the Czech republic 50 years ago and 50 years after. Jindřichův Hradec 09.08.2009 – 11.08.2009. In: IDIMT-2009 System and Humans – A Complex Relationship. Linz : Trauner Verlag universitat, 2009, s. 105–114. ISBN 978-3-85499-624-8.

[4]KAČEROVÁ, E. 2008. International migration and mobility of the EU citizens in the Visegrad group countries: Comparison and bilateral flows. In: European Population Conference. Barcelona. EPC, 142.

[5]KOSCHIN, F. 2005. Kapitoly z ekonomické demografie. Oeconomica, Praha.

[6]LANGHAMROVÁ, JITKA., ARLTOVÁ, M., LANGHAMROVÁ, JANA. 2009. Změny ve věkové struktuře obyvatelstva a jejich možné důsledky. Šlapanice 22.01.2009 – 23.01.2009. In: Bíla místa teorie a černé díry reforem ve veřejném sektoru. Brno : Tribun EU, 2009, s. 26–32. ISBN 978-80-7399-700-7.

[7]http://www.czso.cz/csu/2010edicniplan.nsf/p/4003-10.

Adresa autora (-ov):

Tomáš Fiala, RNDr., Csc.
Katedra demografie, FIaS
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
fiala@vse.cz

Jitka, Langhamrová, doc., Ing., Csc.
Katedra demografie, FIaS
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
langhamj@vse.cz

Zpracování tohoto článku bylo financováno z prostředků institucionální podpory na dlouhodobý koncepční rozvoj vědy výzkumu na FIS VŠE v Praze.

Změny demografického vývoje a populační struktury v ČR od roku 1989
Changes in the demographic development and structure of the population in the Czech Republic since 1989

Tomáš Fiala, Jitka Langhamrová

Abstract: The paper deals with a brief analysis of the changes of the demographic behavior and age structure of the population of the Czech Republic in the past twenty years. It compares the demographic development since 1990 in the Czech Republic with the development up to the end of 1989 and shows the changes in mortality, fertility, migration and also in the age structure.

Abstrakt: Článek uvádí stručnou analýzu změn demografického chování a věkové struktury obyvatelstva České republiky během uplynulých 20 let. Porovnává demografický vývoj od roku 1990 v České republice s demografickým vývojem do konce roku 1989 a poukazuje na rozdíly v úmrtnosti, plodnosti, migraci a také ve věkové struktuře.

Key words: demographic development**,** productive age, population ageing.

Klíčová slova: demografický vývoj, produktivní věk, stárnutí populace.

JEL classification: J11

1. Introduction

As a result of the political, social and economic changes that occurred in the Czech Republic since 1989 there have also been changes in the demographic behavior of the population of the Czech Republic in the past twenty years, some of them quite considerable. These understandably resulted, among other things, in changes in the age structure of the population.

2. Development of the demographic behavior

Mortality

Whereas in the fifties, sixties, seventies and eighties the life expectation in the Czech Republic grew only very slightly or even stagnated or dropped for a short period, from 1990 we have recorded its constant and faster growth, especially in the case of men (see Table 1).

The annual values of the life expectancy are affected of course by random fluctuation. To eliminate these fluctuations we have used the value of the slope regression coefficient to characterize the average annual increase of the life expectation in a given time period. (E. g. the average annual increase of the life expectancy in the period 1950–54 can be expressed by the slope regression coefficient based on linear regression model of life expectancy in the period 1950–54.)

The values of the "moving regression" coefficients for five-years time periods are shown in Figure 1. (The value for each year t is the slope regression coefficient of the linear regression model in the time period t-2; t+2.). The relatively high annual increase in the early fifties has diminished and was followed by period of decrease in the sixties for males and later for a short time also for females. The period of little increase in the early seventies was followed by the stagnation period in late seventies and early eighties. Since late eighties we can observe permanent increase of life expectancy. After 1989 the values of average annual increase of the life expectancy are relatively high and contrary of the previous period higher for males then for females.

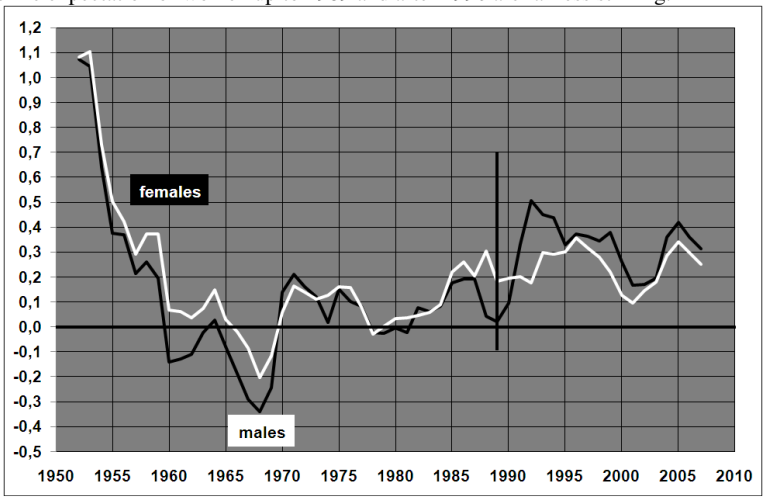
Table 5: Demographic development of the Czech Republic since 1950

Year	Life expectancy		Total fertility rate	Live births	Deaths	Natural increase	Net migration*	Total increase
	males	females						
1950	62,16	66,97	2,828	188 341	103 203	85 138	1 267	86 405
1951	62,16	66,97	2,791	185 570	102 658	82 912	12 406	95 318
1952	64,56	69,41	2,718	180 143	97 726	82 417	8 444	90 861
1953	65,49	70,11	2,615	172 547	98 837	73 710	-1 253	72 457
1954	65,86	70,81	2,578	168 402	99 636	68 766	-14 954	53 812
1955	66,74	71,79	2,578	165 874	93 300	72 574	-9 141	63 433
1956	67,13	72,22	2,562	162 509	93 526	68 983	-6 402	62 581
1957	66,74	71,92	2,478	155 429	98 687	56 742	-5 746	50 996
1958	67,70	72,85	2,280	141 762	93 697	48 065	-6 460	41 605
1959	67,53	72,94	2,092	128 982	97 159	31 823	-4 524	27 299
1960	68,03	73,58	2,091	128 879	93 863	35 016	-6 057	28 959
1961	67,55	73,41	2,112	131 019	94 973	36 046	-6 789	29 258
1962	66,99	72,95	2,120	133 557	104 318	29 239	-5 877	23 363
1963	67,41	73,56	2,310	148 840	100 129	48 711	-3 423	45 289
1964	67,55	73,68	2,338	154 420	101 984	52 436	-6 886	45 551
1965	67,15	73,42	2,178	147 438	105 108	42 330	-8 172	34 159
1966	67,25	73,76	2,022	141 162	105 784	35 378	-9 573	25 806
1967	67,16	73,67	1,908	138 448	108 967	29 481	-14 967	14 515
1968	66,60	73,46	1,838	137 437	115 195	22 242	-13 262	8 981
1969	66,02	73,15	1,874	143 165	120 653	22 512	-14 424	8 089
1970	66,12	73,01	1,928	147 865	123 327	24 538	-16 050	8 489
1971	66,18	73,32	1,994	154 180	122 375	31 805	-1 817	29 988
1972	67,22	73,66	2,094	163 661	119 205	44 456	-1 423	43 033
1973	66,52	73,64	2,312	181 750	124 437	57 313	308	57 621
1974	66,76	73,54	2,462	194 215	126 809	67 406	-1 255	66 151
1975	67,01	73,94	2,432	191 776	124 314	67 462	-1 906	65 556
1976	67,07	74,14	2,390	187 378	125 232	62 146	-1 677	60 469
1977	67,13	74,14	2,339	181 763	126 214	55 549	-3 000	52 549
1978	67,21	74,23	2,326	178 901	127 136	51 765	-2 243	49 522
1979	67,35	74,30	2,272	172 112	127 949	44 163	-1 813	42 350
1980	66,84	73,92	2,072	153 801	135 537	18 264	-2 451	15 813
1981	67,18	74,30	1,996	144 438	130 407	14 031	-4 235	9 796
1982	67,27	74,40	2,002	141 738	130 765	10 973	-4 204	6 769
1983	67,02	74,25	1,968	137 431	134 474	2 957	-3 569	-612
1984	67,31	74,18	1,968	136 941	132 188	4 753	-3 331	1 422
1985	67,46	74,70	1,952	135 881	131 641	4 240	-3 757	483
1986	67,46	74,62	1,936	133 356	132 585	771	-2 939	-2 168
1987	67,83	75,13	1,910	130 921	127 244	3 677	-3 231	446
1988	68,09	75,27	1,940	132 667	125 694	6 973	-3 408	3 565
1989	68,11	75,40	1,874	128 356	127 747	609	-4 493	-3 884
1990	67,54	76,01	1,893	130 564	129 166	1 398	-5 328	-3 930
1991	68,21	75,67	1,861	129 354	124 290	5 064	-576	4 488
1992	68,52	76,11	1,715	121 705	120 337	1 368	8 329	9 697
1993	69,28	76,35	1,666	121 025	118 185	2 840	2 024	4 864
1994	69,53	76,55	1,438	106 579	117 373	-10 794	6 490	-4 304
1995	69,96	76,94	1,278	96 097	117 913	-21 816	6 547	-15 269
1996	70,37	77,27	1,185	90 446	112 782	-22 336	6 677	-15 659
1997	70,50	77,49	1,173	90 657	112 744	-22 087	8 623	-13 464
1998	71,13	78,06	1,157	90 535	109 527	-18 992	6 036	-12 956
1999	71,40	78,13	1,133	89 471	109 768	-20 297	5 322	-14 975
2000	71,65	78,35	1,144	90 910	109 001	-18 091	3 087	-15 004
2001	72,14	78,45	1,146	90 715	107 755	-17 040	-8 551	-25 591
2002	72,07	78,54	1,171	92 786	108 243	-15 457	12 290	-3 167
2003	72,03	78,51	1,178	93 685	111 288	-17 603	25 789	8 186
2004	72,55	79,04	1,226	97 664	107 177	-9 513	18 635	9 122
2005	72,88	79,10	1,282	102 211	107 938	-5 727	36 229	30 502
2006	73,45	79,67	1,328	105 831	104 441	1 390	34 720	36 110
2007	73,67	79,90	1,438	114 632	104 636	9 996	83 945	93 941
2008	73,96	80,13	1,500	119 570	104 948	14 622	71 790	86 412
2009	74,19	80,13	1,490	118 348	107 421	10 927	28 344	39 271

*In 19502000 including unregistered migration estimated by the census.

Source: Czech Statistical Office.

The application of the linear regression model for the period 1969–1989 shows that the life expectancy of males increased annually only 0.08 years, the annual increase for females was 0.10. From 1989 up to the present its growth for males has been more than fourfold – it rose annually by 0.33 years (i. e. four months). For females in the period 1989-2009 the annual increase was "only" 0.24 years (i. e. three months). The differences in the development of the life expectation of women up to 1989 and after 1990 are far less striking.



Source: Own computations based on the data of the Czech Statistical Office

Figure 1: Average annual increase of the life expectancy (slope regression coefficient of a five-year moving linear regression)

Fertility

Probably few demographers expected such a marked decline in fertility as that which took place immediately in the first years after 1989. In the period 1995–2005 the values of total fertility rate were lower than 1.3 (see Table 1) and fertility in the Czech Republic was among the lowest in the EU at that time. The later relatively rapid growth was caused chiefly by the realization of the deferred fertility of women who had decided not to have children until around the age of 30, which is the average age of mothers in the countries of western, northern and southern Europe. In 2009 the value of overall fertility was somewhat lower than in the previous year and the data for the first half of 2010 indicate that the stagnation in fertility in the Czech Republic might proceed.

Live births, deaths, natural increase

The drop in the number of live births, which began in 1975, continued also after 1989, but at a far more rapid rate (see Table 1). Whereas in 1991 there were still almost 130,000 live births in the Czech Republic, in the years 1996–2001 the annual number of live births was only somewhere around 90,000; this means more than 30 % lower than at the beginning of the nineties. In the first years of this millennium, thanks to the strong population generations of women born in the seventies, the number of live births gradually increased to almost 120,000. In the next few years, however, due to the constantly falling numbers of women at the age of greatest fertility, a long-term decline is again to be expected.

The development of the number of deaths was relatively stable and after stagnation in the eighties it began to drop slightly from 1990.

From the end of the First World War the Czech Republic constantly had a positive natural increase, the values of which reached about several tens of thousands of people each year. From 1983, however, these dropped below 10,000 persons. With regard to the above-mentioned sharp decline in the number of live births, the Czech Republic had in the period 1994–2005, for the first time in its history, a negative value of the natural increase, which was not caused by any catastrophic events, but merely by a deep decline in fertility. (See the Table 1.) The present situation, where the number of live births is again higher than the number of deaths, will probably end in a few years and the natural increase of the population of the Czech Republic will be again long-term (if not permanently) negative. Only a constantly positive and gradually increasing net migration will then prevent a decrease of the population.

Migration

Of all the demographic processes it is usually migration that is most influenced by political, social and economic changes taking place on the territory of the country studied. This is logical; as opposed to mortality and fertility migration may be (and usually is) directly or indirectly regulated by the laws of the country concerned. In the time of the socialist regime in Czechoslovakia the possibility of migration from or to the Czech Republic (with the exception of migration from/to Slovakia) was strongly restricted. If we add illegal emigration to the net migration published by the Czech Statistical Office, we can then say that (in spite of the constantly positive net migration with Slovakia) up to the end of 1989 (with the exception of the beginning of the fifties) the Czech Republic had a negative net migration deficit (see the Table 1).

Since 1992 this trend changes – the annual number of immigrants is (with the only exception in the year 2001) higher than the number of emigrants. The positive net migration has partly eliminated the natural decrease of the population of the Czech Republic. Since 2002 the values of the annual net migration balance have been about tens of thousands. In 2009, however, there was a relatively dramatic drop in the net migration balance from a value of more than 70,000 to less than 30,000. The data for the 1st half of 2010 indicate a further marked drop in the net migration; its value was at the level of only about 16 % of the net migration in the 1st half of 2009. (Data since 2001 are not corrected by unregistered migration, it will be possible to make the correction retrospectively when the results of the census in 2011 will be available.)

3. The development of the age structure

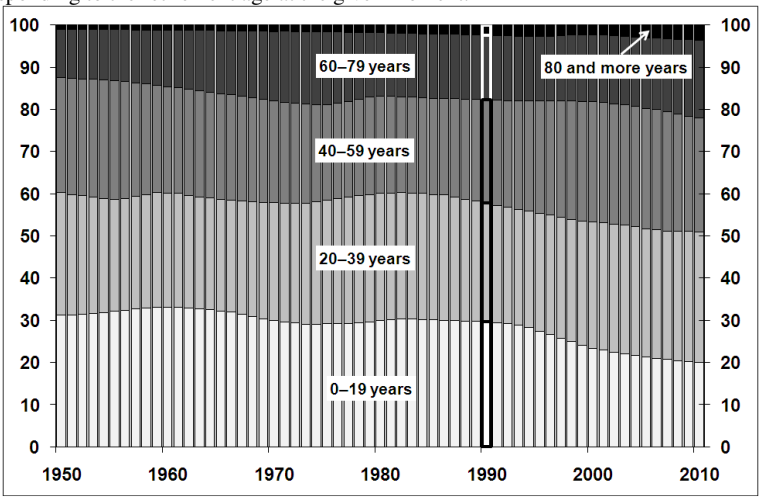
The above changes in demographic development naturally resulted in changes in the age structure of the population of the Czech Republic. These changes also, however, naturally took place smoothly and slowly.

The ageing of the population continued. There was a strong reduction in particular of the proportion of people under the age of 20. Whereas at the beginning of 1990 this proportion was still around 30 %, in the course of the next 10 years it dropped to less than 25 % and at present it is only just over 20 %. The proportion of persons over the age of 60, on the contrary, increased from just below 18 % (1990) to more than 22 % (2010). The proportions of those aged 20 – 39 years and 40 – 59 years also increased by a few percentage points (see the Fig. 2).

In relation to the ageing of the population there is often talk of the increase in the financial burden of the current pension system. A simple measure of this burden is what is known as the old-age dependency ratio – the ratio of the number of persons of pension age to the number of persons of productive age. A more suitable age for the lower limit of the productive age is not the 15 years used formerly, but as much as 20 years. In the present day only very few young people begin economic activity immediately after leaving elementary

school – usually they raise their qualifications further at secondary school or in an apprenticeship course.

The upper limit of the productive age is usually considered to be the normal age for retirement. Because the retirement age for both men and women has been rising gradually in the Czech Republic since 1996, the upper limit of the productive age was set as variable – corresponding to the retirement age at the given moment.



Source: Own computations based on the data of the Czech Statistical Office

Figure 2: Development of the age structure of the population of the Czech Republic (in %)

The retirement age for women in the Czech Republic depends on the number of children reared, so for the sake of simplicity it was assumed that each woman reared just 2 children.

Up to the end of 1995 the retirement age for men was taken to be 60 years and for women 55 years. Since 1996 there has been an increase in the retirement age. The increase is carried out on a generation basis, not as a transversal; for each further year of birth the retirement age for men rises by 2 months and in the case of women it is 4 months more than for the previous year of birth. It is clear that from the transversal viewpoint the retirement age for men rises by 1 year in 7 years. (At the end of 1995 the retirement age for men was 60 for men born in 1935. Men born in 1941 had retirement age 61 years, they reached the retirement age in 2002, so at the end of 2002 the retirement age was 61 years). Similarly the retirement age for women rises by 1 year in the course of 4 years.

Perhaps somewhat surprising is the finding that the highest values of the old-age dependency ratio (around 44) were achieved at the beginning of the seventies. By the end of the seventies it had dropped to a value just over 40 and it remained at these values up to the beginning of the nineties. From 1994 the value of this index has constantly remained lower than 40. With regard to the raising of the retirement age the drop in the values of this index continued down to 36 and at present it is slightly over 37. (If, however, the retirement age had remained without change up to the present at the level of the beginning of the nineties, the present value of the index would already be over 47.) See the Table 2. It seems, then, that the raising of the retirement age was relatively well timed and its rate appears to be optimal.

The old-age dependency ratio is, of course, a very rough index. It does not take into account whether persons of productive age really are economically active and whether people of post-productive age really are receiving a pension. Simultaneously with the raising of

retirement age it was made possible to retire prematurely. In spite of this it appears that to ascribe the present problems with financing the pension insurance system exclusively to unfavorable demographic development is simplifying matters or is even misleading.

Table 2: Development of the old-age dependency ratio in the Czech Republic

Year	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970
OADR	29,1	29,7	30,3	31,0	31,7	32,7	33,7	34,7	35,7	36,8	37,3	38,4	39,4	40,3	41,2	42,0	42,6	43,1	43,5	44,0
Year	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
OADR	44,2	43,5	42,8	41,9	41,7	41,4	41,0	40,6	40,2	40,3	40,5	40,9	41,1	41,2	41,2	41,2	41,2	41,1	41,1	40,9
Year	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
OADR	41,0	40,7	40,3	39,8	39,3	38,8	38,0	37,3	36,7	36,3	36,0	36,0	36,0	36,1	36,5	36,7	36,9	36,9	37,1	37,6

Source: Own computations based on the data of the Czech Statistical Office

4. Conclusions

One of the most marked changes in the demographic behavior of the population of the Czech Republic after 1989 was the sharp drop in fertility, which resulted in the natural decline in the population of the Czech Republic in the 19942005 period. As opposed to the preceding period immigration prevailed over emigration from 1990. Especially after the year 2000 the net migration (which partly compensated for the drop in population through natural change) achieved values in the order of several tens of thousands per annum.

Thanks to the relatively well-timed raising of the retirement age and the optimally selected rate of its growth the ratio of the number of persons of retirement age to the number of persons of productive age is so far relatively stable in spite of the ageing of the population.

5. References

[1]FIALA, T. 2008. Analýza růstu střední délky života v ČR metodou klouzavé lineární regrese. Forum Statisticum Slovacum [CD-ROM], roč. 6, č. 6, s. 31–35. ISSN 1336-7420.

[2]FIALA, T., LANGHAMROVÁ, J. 2010. Ekonomická aktivita mužů a žen od konce roku 2000 a vliv zvyšování důchodového věku. Demografie [CD-ROM], roč. 52, č. 1, s. 110–122. ISSN 0011-8265.

[3]FIALA, T., LANGHAMROV, JITKA, LANGHAMROVÁ, JANA. 2009. Changes in the Age Structure of the Population in the Czech Republic and its Economics Consequences. Uherské Hradiště 27.08.2009 – 28.08.2009. In: AMSE 2009 Applications of Mathematics and Statistics in Economy. [online] Praha : Oeconomica, s. 137–148. ISBN 978-80-245-1600-4. URL: http://amse2009.vse.cz/proceedings.pdf.

[4]Zákon o důchodovém pojištění 155/1995 Sb. Sbírka zákonů ČR

[5]hhttp://www.czso.cz/

Adresa autorů:

Tomáš Fiala, RNDr., Csc.
katedra demografie
fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
fiala@vse.cz

Jitka, Langhamrová, doc., Ing., Csc.
katedra demografie
fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
langhamj@vse.cz

The article originated within the framework of the long-term research project 2D06026 "The Reproduction of Human Capital", financed by the Ministry of Education, Youth and Sport within the National Research Programme II.

Výpočet variability pre nelineárne odhady
Variance computing for nonlinear estimates

Boris Frankovič

Abstract: This article is about computing variance for nonlinear estimates used in practice in statistical offices. We'll express this variance using Taylor approximation to reduce non-linear forms of the variables to linear forms and using jackknife method too. In both cases we'll deal with stratified sample.

Key words: variance, nonlinear, jackknife, stratified

Kľúčové slová: variabilita, disperzia, nelineárne, jackknife, stratifikovaný

JEL classification: C14, C83

1. Úvod

Disperzia je kľúčovým ukazovateľom presnosti odhadov pri hodnotení kvality štatistických výstupov vybraných štatistických zisťovaní. V tomto článku sa budeme špeciálne zaoberať nelineárnymi odhadmi, ktoré sa často vyskytujú v štatistickej praxi. Výpočet podľa štandardného vzorca na výpočet disperzie nie je možný. Pristúpime k aproximačným metódam, ako ku klasickej linearizačnej metóde, tak aj k metóde jackknife a v oboch situáciách sa zameriame na výpočet v prípade stratifikovaného výberu. Výsledky naprogramujeme v štatistickom softwari R.

2. Linearizačná metóda

Často používaná metóda na výpočet variancie komplikovaného odhadu je založená na Taylorovej aproximácii, kedy nelineárny tvar odhadu prevedieme na tvar lineárny. Označme tento nelineárny odhad ako F. F je funkcia m premenných (6, s. 411). Pre dostatočne veľkú výberovú vzorku nám postačia prvé dva členy Taylorovho rozvoja.

$$f(x) = f(x_0) + \left.\frac{\partial f(x)}{\partial x}\right|_{x_0} (x - x_0) + O(h)$$

kde $O(h)$ ide v limite do nuly. Aplikovaním na výpočet disperzie dostaneme

$$D\left(F(X_1, \dots, X_m)\right) = E\left[F(\vec{X}) - E\left(F(\vec{X})\right)\right]^2 =$$

$$= E\left[F(\overrightarrow{X_0}) + \left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} (\vec{X} - \overrightarrow{X_0}) - E\left(F(\overrightarrow{X_0}) + \left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} (\vec{X} - \overrightarrow{X_0})\right)\right]^2 =$$

$$= E\left[\left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} (\vec{X} - \overrightarrow{X_0}) - E\left(\left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} (\vec{X} - \overrightarrow{X_0})\right)\right]^2$$

Ďalej

$$E\left(\left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} (\vec{X} - \overrightarrow{X_0})\right) = \left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} E(\vec{X} - \overrightarrow{X_0}) = \left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} E(\vec{X}) - \left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|^T_{\overrightarrow{X_0}} E(\overrightarrow{X_0})$$

a tento výraz sa rovná nule vtedy a len vtedy, keď $E(\vec{X}) = E(\overrightarrow{X_0})$, teda keď $E(\vec{X}) = \overrightarrow{X_0}$.
Preto budeme parciálne derivácie robiť v bodoch, ktoré zodpovedajú stredným hodnotám jednotlivých premenných. Dostávame teda

$$D\big(F(X_1,\dots,X_m)\big) = E\left[\left.\frac{\partial F(\vec{X})}{\partial \vec{X}}\right|_{E(\vec{X})}^{T}\left(\vec{X}-E(\vec{X})\right)\right]^2 =$$

$$E\left[\sum_{i=1}^{m}\left.\frac{\partial F(\vec{X})}{\partial X_i}\right|_{E(\vec{X})}\big(X_i-E(X_i)\big)\right]^2 = D\left(\sum_{i=1}^{m}\left.\frac{\partial F(\vec{X})}{\partial X_i}\right|_{E(\vec{X})}X_i\right) \qquad (1)$$

Výsledný vzťah je uvedený v (6, s. 411). V príklade, na ktorom sme skúšali túto teóriu, sa vyskytoval odhad typu

$$F = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}} \qquad (1, \text{s. } 17) \qquad (2)$$

pričom

H je počet nenulových strát

h je číslo straty

n_h je počet jednotiek v strate h

n je počet jednotiek vo výbere, $n = \sum_{h=1}^{H} n_h$

w_{hi} je váha i-tej jednotky v h-tej strate

y_{hi}, x_{hi} sú hodnoty analyzovaných premenných i-tej jednotky v h-tej strate, či už kategorické, alebo numerické

Označme $w_{hi}y_{hi}, w_{hi}x_{hi}$ ako X_{khi}, $k = 1{,}2$. Dostaneme

$$F = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{1hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}}$$

(6, s. 412). F je funkcia $m = 2n$ premenných, spočítame teda obe parciálne derivácie v príslušných bodoch. Stredné hodnoty jednotlivých argumentov funkcie sa budú rovnať ich hodnotám, nakoľko nie sú náhodnými premennými.

$$\left.\frac{\partial F(\vec{X})}{\partial X_{1hi}}\right|_{E(\vec{X})} = \frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}} \qquad \left.\frac{\partial F(\vec{X})}{\partial X_{2hi}}\right|_{E(\vec{X})} = -\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}\right)^2}$$

Dosadením do vzťahu (1)

$$D\big(F(X_1,\dots,X_m)\big) = D\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}}X_{1hi} - \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}\right)^2}X_{2hi}\right) =$$

$$= D\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}}w_{hi}y_{hi} - \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}\right)^2}w_{hi}x_{hi}\right)$$

Počítajme

$$\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}}w_{hi}y_{hi} - \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} X_{2hi}\right)^2}w_{hi}x_{hi} =$$

$$= \frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}}w_{hi}y_{hi} - \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}y_{hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}\right)^2}w_{hi}x_{hi} =$$

$$= \frac{w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}} - \frac{w_{hi}x_{hi}F}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}} = \frac{w_{hi}(y_{hi}-x_{hi}F)}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}} := g_{hi} \qquad (1, \text{s. } 18)$$

Pokračujme

$$D\big(F(X_1,\dots,X_m)\big)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}D(g_{hi})=\sum_{h=1}^{H}\sum_{i=1}^{n_h}(1-\pi_{hi})\frac{1}{N_h-1}\sum_{i=1}^{N_h}(g_{hi}-\bar{g}_h)^2$$

(6, s. 413) kde N_h je počet jednotiek v opore v strate h, $\bar{g}_h=\frac{\sum_{i=1}^{N_h}g_{hi}}{N_h}$ (1, s.18) je priemer hodnôt g_{hi} vo vnútri straty h a $\pi_{hi}=\frac{1}{w_{hi}}$. $(1-\pi_{hi})$ je korekcia disperzie pri výbere bez opakovania (4, s. 5225), pri výbere s opakovaním (čo nie je teraz náš prípad) sa táto korekcia nepoužíva. Štandardne sa využíva korekcia $\left(1-\frac{n_h}{N_h}\right)$, avšak vhodná je iba v prípade, že všetky váhy v rámci straty sú rovnaké. V opačnom prípade je prijateľnejšia práve korekcia $(1-\pi_{hi})$, ktorá zohľadňuje váhu každej jednotky zvlášť.

$$\widehat{D}\big(F(X_1,\dots,X_m)\big)=\sum_{h=1}^{H}\sum_{i=1}^{n_h}(1-\pi_{hi})\frac{1}{n_h-1}\sum_{i=1}^{n_h}(g_{hi}-\bar{g}_h)^2=$$

$$=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(1-\pi_{hi})(g_{hi}-\bar{g}_h)^2\quad(1,\text{s. }18)\qquad(3)$$

$$\bar{g}_h=\frac{\sum_{i=1}^{n_h}g_{hi}}{n_h}\quad(1,\text{s. }18)$$

Ak $n_h=1$ pre nejaké h, tak disperzia vo vnútri tejto straty bude nulová. Sumujeme teda len cez straty, kde sú aspoň dve jednotky.

Takisto často sa vyskytujúci ukazovateľ je jednoduchšieho tvaru

$$F'=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}}\quad(1,\text{s. }16)\qquad(4)$$

Označme $w_{hi}y_{hi}$, w_{hi} ako X_{khi}, $k=1{,}2$. Pokračujme analogicky s predošlým postupom, dostaneme

$$D\big(F'(X_1,\dots,X_m)\big)=D\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{2hi}}w_{hi}y_{hi}-\sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{2hi}\right)^2}w_{hi}\right)$$

$$\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{2hi}}w_{hi}y_{hi}-\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{1hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}X_{2hi}\right)^2}w_{hi}=$$

$$=\frac{1}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}}w_{hi}y_{hi}-\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}y_{hi}}{\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}\right)^2}w_{hi}=$$

$$=\frac{w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}}-\frac{w_{hi}F'}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}}=\frac{w_{hi}(y_{hi}-F')}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}w_{hi}}:=e_{hi}\quad(1,\text{s. }16)$$

$$\bar{e}_h=\frac{\sum_{i=1}^{n_h}e_{hi}}{n_h}$$

$$\widehat{D}\big(F'(X_1,\dots,X_m)\big)=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(1-\pi_{hi})(e_{hi}-\bar{e}_h)^2\quad(1,\text{s. }16)\qquad(5)$$

3. Metóda jackknife

Táto metóda je jednou z viacerých replikačných metód. Bola navrhnutá M. H. Quenouilleom v roku 1949 za účelom zostrojenia univerzálneho odhadu odchýlky od

skutočných hodnôt parametrov (7, s. 4). Stavebným kameňom sú replikované odhady ukazovateľa vypočítaného z $n-1$ (v prípade stratifikácie n_h-1 v každej strate zvlášť, vo všeobecnom prípade $n-d$, $d\epsilon\mathbb{N}$) pozorovaní.

Nech $\hat{\theta}$ je odhad ukazovateľa na základe celého výberového súboru. Ďalej nech $\hat{\theta}_{n_h-1}(i)$ je odhad rovnakého funkčného tvaru ako $\hat{\theta}$, avšak vypočítaný po vynechaní i-tej jednotky v h-tej strate (3, s. 28), pričom zvyšných n_h-1 jednotiek bude prevážených výrazom $\frac{n_h}{n_h-1}$ ako kompenzácia za jednu vynechanú (4, s. 5225). Dostaneme teda n_h replikovaných odhadov v každej strate, dokopy n.

Najskôr sa zameriame na jackknife odhad v prípade jednoduchého náhodného výberu a potom odvodené výsledky aplikujeme na stratifikovaný výber. Pseudohodnoty sú definované ako

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{n-1}(i)$$

(3, s. 28) a jackknife odhad ukazovateľa je priemer pseudohodnôt, čiže podľa (5, s. 8)

$$\hat{\theta}_{JK} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \frac{1}{n}\sum_{i=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{n-1}(i)\right) =$$

$$= n\hat{\theta} - \frac{n-1}{n}\sum_{i=1}^{n}\hat{\theta}_{n-1}(i) = n\hat{\theta} - (n-1)\hat{\theta}_{\bullet} \qquad (6)$$

kde $\hat{\theta}_{\bullet} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{n-1}(i)$ je aritmetický priemer pseudohodnôt (7, s. 5). Uvedomiac si, že jackknife odhad je priemer, dostávame jeho disperziu vyjadrenú na základe (5, s. 8) v tvare

$$\hat{D}\left(\hat{\theta}_{JK}\right) = \frac{\hat{D}\left(\hat{\theta}_i\right)}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\theta}_i - E\left(\hat{\theta}_i\right)\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\hat{\theta}_i - \hat{\theta}_{JK}\right)^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{n-1}(i) - n\hat{\theta} + (n-1)\hat{\theta}_{\bullet}\right)^2 =$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[-(n-1)\hat{\theta}_{n-1}(i) + (n-1)\hat{\theta}_{\bullet}\right]^2 =$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[(n-1)\left(\hat{\theta}_{\bullet} - \hat{\theta}_{n-1}(i)\right)\right]^2 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{\bullet} - \hat{\theta}_{n-1}(i)\right)^2 =$$

$$= \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{n-1}(i) - \hat{\theta}_{\bullet}\right)^2$$

Výsledný vzťah je uvedený v (7, s. 5). Po prenásobení korekciou

$$\hat{D}\left(\hat{\theta}_{JK}\right) = \frac{n-1}{n}\sum_{i=1}^{n}(1-\pi_i)\left(\hat{\theta}_{n-1}(i) - \hat{\theta}_{\bullet}\right)^2 \qquad (7)$$

Tento vzorec je však použiteľný iba v situáciách, keď sa jedná o odhad lineárneho ukazovateľa v prípade jednoduchého i stratifikovaného výberu a o odhad zložitého nelineárneho ukazovateľa v prípade jednoduchého náhodného výberu. Náš ukazovateľ je nelineárny odhad v prípade stratifikovaného výberu, môžeme teda buď použiť klasickú linearizačnú metódu, alebo metódu jackknife ďalej linearizovať, čoho výsledkom budú identické výsledky ako pri klasickej linearizácii.

Nakoľko je vzorcom (7) vypočítateľná disperzia odhadu úhrnu aj v prípade stratifikovaného výberu, označme odhad tretieho ukazovateľa ako podiel dvoch úhrnov. F je rovnako ako v (2)

$$F = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}}$$

Označme jednotlivé úhrny.

$$F = \frac{Y}{X} \tag{8}$$

V (7) sme vyjadrili tvar jackknife disperzie bez stratifikácie, z čoho vyplýva, že disperzia celkového úhrnu bude

$$\widehat{D}(Y) = \widehat{D}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}y_{hi}\right) = \sum_{h=1}^{H}\widehat{D}\left(\sum_{i=1}^{n_h} w_{hi}y_{hi}\right) = \sum_{h=1}^{H}\widehat{D}\left(\hat{\theta}_{JKh}\right)$$

kde $\hat{\theta}_{JKh}$ je jackknife odhad úhrnu v strate h. Čiže

$$\widehat{D}(Y) = \sum_{h=1}^{H}\frac{n_h - 1}{n_h}\sum_{i=1}^{n}(1-\pi_{hi})\left(\hat{\theta}_{n_h-1}(i) - \hat{\theta}_{h\bullet}\right)^2 \tag{9}$$

(4, s. 5225), pričom značenie je analogické s predošlým.

Linearizujme teraz disperziu odhadu rovného podielu dvoch úhrnov. Postupujme ako vo vzorci (1).

$$D(F) = D\left(\frac{Y}{X}\right) = E\left[\left.\frac{\partial F(Y,X)}{\partial (Y,X)^T}\right|_{E(X,Y)}^{T}\left(\begin{pmatrix}Y\\X\end{pmatrix} - \begin{pmatrix}E(Y)\\E(X)\end{pmatrix}\right)\right]^2 =$$

$$= E\left[\left.\frac{\partial F}{\partial Y}\right|_{E(X,Y)}\big(Y - E(Y)\big) + \left.\frac{\partial F}{\partial X}\right|_{E(X,Y)}\big(X - E(X)\big)\right]^2 =$$

$$= E\left[\frac{1}{E(X)}\big(Y - E(Y)\big) - \frac{E(Y)}{E^2(X)}\big(X - E(X)\big)\right]^2 =$$

$$= E\left[\frac{1}{E(X)}\left(Y - E(Y) - \frac{E(Y)}{E(X)}\big(X - E(X)\big)\right)\right]^2$$

Keďže presnú hodnotu $E(X)$, resp. $E(Y)$ nepoznáme, spočítame aritmetické priemery.

$$\hat{E}(X) = \hat{E}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi}\right) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\hat{E}(w_{hi}x_{hi}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi} =$$

$$= n\frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{n_h} w_{hi}x_{hi} = Y$$

Pokračujeme

$$\widehat{D}(F) = \frac{1}{E^2(X)}E\left[\big(Y - E(Y)\big)^2 - 2\big(Y - E(Y)\big)\frac{E(Y)}{E(X)}\big(X - E(X)\big) + \left(\frac{E(Y)}{E(X)}\right)^2\big(X - E(X)\big)^2\right]$$

$$= \frac{1}{X^2}\left[\widehat{D}(Y) + \left(\frac{E(Y)}{E(X)}\right)^2\widehat{D}(X) - 2\left(\frac{E(Y)}{E(X)}\right)\mathrm{cov}(Y,X)\right] =$$

$$= \frac{1}{X^2}\left[\widehat{D}(Y) + F^2\widehat{D}(X) - 2F\mathrm{cov}(Y,X)\right] \tag{10}$$

ako je uvedené v (4, s. 5226). Vo výraze (9) už vystupujú iba variancie a kovariancia lineárnych stratifikovaných odhadov, ktoré sa dajú vypočítať jednak klasickou linearizáciou a jednak použitím vzorca (9).

4. Záver

V tejto práci sme sa zaoberali výpočtom disperzie pre nelineárne ukazovatele, ktoré sa vyskytujú v praxi na štatistických úradoch v rôznych zisťovaniach. Nakoľko klasický výpočet nie je možný, pristúpili sme k aproximačným vyjadreniam, ktoré priniesli uspokojivé

výsledky. Celá procedúra bola následne naprogramovaná v štatistickom softwari R. Nakoľko je algoritmus plne prispôsobený potrebám konkrétneho zisťovania, vstupný súbor je iba dátová tabuľka a výstupný súbor je tabuľka s konkrétnymi hodnotami odhadov, ich smerodajných odchýlok a variačných koeficientov. Nebolo by však problémom tento algoritmus zovšeobecniť, telo funkcie by pocítilo iba malé zmeny.

5. Literatúra

[1] JURIOVÁ, J.: Ukazovatele kvality pre vybrané štatistické zisťovania. Bratislava: Infostat, 2008, s. 16 – 18.

[2] KREWSKI, D. – RAO, J.N.K.: Large sample properties of the linearization, jackknife and balanced repeated replication methods. Health & Welfare Canada and Carleton University, s. 457 - 459 http://www.amstat.org/sections/srms/proceedings/papers/1978_093.pdf (2010-05-05)

[3] RAMASUBRAMANIAN, V. – RAI, A. – SINGH, R.: Jackknife variance estimation under two-phase sampling. New Delhi: Indian Agricultural Statistics Research Institute, s. 28. http://iospress.metapress.com/content/r21k217h54175360/fulltext.pdf (2010-03-29)

[4] STEEL, P. – McNERNEY, V. – SLANTA, J.: An Investigation of Stratified Jackknife Estimators Using Simulated Establishment Data Under an Unequal Probability Sample Design. Section on Survey Research Methods – JSM, 2009, s. 5225 – 5226. http://www.amstat.org/sections/srms/Proceedings/y2009/Files/305696.pdf (2010-05-10)

[5] WALSH, B.: Resampling Methods: Randomization Tests, Jackknife and Bootstrap Estimators. 2000, s. 8. http://nitro.biosci.arizona.edu/courses/EEB581-2006/handouts/random.pdf (2010-03-31)

[6] WOODRUFF, R.S.: A Simple Method for Approximating the Variance of a Complicated Estimate. Davis: Journal of the American Statistical Association, 1971, s. 411 – 413.

[7] Descriptive Statistics for Research: Lecture 6: Resampling methods. University of Oxford: Institute for the Advancement of University Learning & Department of Statistics. 2002, s. 4 – 5. http://www.stats.ox.ac.uk/pub/bdr/IAUL/Course1Notes6.pdf (2010-03-29)

Adresa autora:

Boris Frankovič, Mgr.
Pod Párovcami 113
921 01 Piešťany
Boris.Frankovic@statistics.sk

Modely hodnotenia spokojnosti zamestnancov
Evaluation models of employee satisfaction

Marek Andrejkovič, Zuzana Hajduová

Abstract: The issue of employee satisfaction is very important for businesses. In this article we propose a model of evaluation of employee satisfaction. Using of this model is dependent on using of computer technology because the amount of calculations needs to be done in iterations. The usage of this model is possible in various fields. The usage in field of evaluation of employee satisfaction is only illustrative.

Key words: Employee Satisfaction. Information System. Evaluation.

Kľúčové slová: Spokojnosť zamestnancov. Informačný systém. Hodnotenie.

1. Úvod

Meranie spokojnosti zamestnancov s nadriadenými je pre súčasné manažmentom riadené procesy veľmi dôležité. Výkonnosť zamestnancov je podmieňovaná spokojnosťou zamestnancov či už pozitívne alebo negatívne. Preto je potrebné v čase poznať aktuálnu mieru spokojnosti zamestnancov a na základe toho voliť efektívne kroky. Preto v tomto príspevku poukážeme na možnosti využitia nového modelu hodnotenia spokojnosti zamestnancov a pričom využitie informačných technológií je pre tento model nutné. Poukážeme na zložitosť modelu, ktorý je možné odstrániť využitím výpočtovej techniky.

2. Spokojnosť zamestnancov

Pojem spokojnosť predstavuje mieru osobného prežívania jedinca alebo stav určitého naplnenia vyplývajúci z úspechu. Zúžením tohto pojmu môžeme definovať pojem pracovnej spokojnosti ako pocit uspokojenia, ktorý jedincovi prináša je práca, resp. pracovná činnosť. (Kollárik, 2002) V teórii sa môžeme stretnúť s rozličnými hľadiskami, s akými autori nazerajú na tento pojem. Tento pojem môžeme rozčleniť do dvoch hlavných kategórií, a tými sú spokojnosť v práci a spokojnosť s prácou. Spokojnosť v práci je pritom širší pojem a predstavuje okrem spokojnosti s prácou aj spokojnosť s pracovnými podmienkami, s medziľudskými vzťahmi a podobne. (Kollárik, 2002)

V oblasti vedenia zamestnancov sa autori orientujú na spokojnosť zamestnancov so štýlom vedenia nadriadeného. Autori poukazujú na rozličný štýl vedenia, ktorý nemusí vyhovovať všetkým zamestnancom. Pritom každého zamestnanca motivuje k vyšším pracovným výkonom iný prístup vedúceho pracovníka. Z toho dôvodu je potrebné v záujem zvyšovania výkonnosti a tiež dobrej pracovnej klímy odporučiť vedúcemu zamestnancovi vhodný štýl. (Carter et al., 2004) Spokojnosť zamestnancov s oblasťou motivácie zahŕňajú rozličné faktory od podporných nástrojov a vhodnosťou ich použitia a výhodnosti. V dnešnej dobe sa spoločnosti snažia motivovať zamestnancov nefinančnými prostriedkami. V niektorých prípadoch autori uvádzajú, že využitie finančných prostriedkov má omnoho nižšiu efektívnosť ako nefinančné. Je to hlavne v prípade, ak zamestnanci musia podávať vysoký výkon a sú veľmi dobre finančne hodnotený. V takom prípade ďalšie finančné prostriedky majú za následok iba malú mieru motivácie. (Farnham, 2000; Fuchsová – Kravčáková, 2004)

3. Model hodnotenia spokojnosti

Navrhovaný model hodnotenia zamestnancov vychádza z princípu teórie grafov, konkrétne teórie nábojov. Grafické zobrazenie vzťahov medzi hodnotenými je možné realizovať pomocou bipartitného grafu. V najjednoduchšom prípade uvažujme, že všetci zamestnanci hodnotia všetkých nadriadených, pričom ich počet môže ale nemusí byť rovnaký. Uvedený predpoklad neskôr bude zavedený do tohto modelu. Prostredníctvom pridelenia preferencie každého hodnotiaceho ako aj hodnoteného bude možné iterovať uvedený postup relatívne do nekonečna, až kým nedôjde k identifikácii mienkotvorných skupín, ktoré potom určia konečnú preferenciu. Môžeme predpokladať, že v niektorých prípadoch pritom takáto mienkotvorná skupina nebude existovať a vtedy je možné použiť iba konvenčný spôsob zisťovania hodnotenia zamestnancov.

V prípade tohto modelu uvažujme graf *G*, ktorý pozostáva z množiny hrán a množiny vrcholov, čo zapíšeme nasledovne:

$$G=(V;E) \tag{1}$$

kde *V* označuje množinu vrcholov a *E* označuje množinu hrán.

Uvedený graf predstavuje schématické zobrazenie vzťahov medzi subjektmi modelu, teda zamestnancami, ktorí hodnotia nadriadených. To môžeme zobraziť nasledovne ako je to uvedené v obrázku.

Nie všetci zamestnanci musia hodnotiť každého nadriadeného. Taktiež určovanie váh bude predmetom ďalšej podkapitoly. Na základe uvedenej definície grafu teda existujú dve skupiny vrcholov, pričom väzby (hrany) medzi vrcholmi rovnakej skupiny neexistujú.

Vrcholy *V* grafu *G* predstavujú hodnotiteľov aj hodnotených. Tieto dve skupiny môžeme odlíšiť nasledovne. Množinu všetkých vrcholov rozlišujeme na dve podmnožiny, a to podmnožina *A*, ktorá predstavuje hodnotiteľov (zamestnanci) a podmnožiny *B*, ktorá predstavuje hodnotených (nadriadených). Tieto podmnožiny sú disjunktné, to znamená, že prienikom týchto dvoch podmnožín je prázdna množina. Teda platí, že

$$A \cap B = \varnothing \tag{2}$$

Pritom podmnožina hodnotiteľov pozostáva z nasledujúcich prvkov $A=\{a_1,a_2,\ldots,a_m\}$. Počet hodnotiteľov je teda *m*. Druhá podmnožina hodnotených pozostáva z nasledujúcich prvkov, a to $B=\{b_1,b_2,\ldots,b_n\}$. Spolu to zapíšeme ako

$$V = A \cup B \tag{3}$$

Množina hrán *E* pozostáva z nasledujúcich častí $E=\{e_1,e_2,\ldots,e_{m\cdot n}\}$. Hrany vedú z každého vrcholu z podmnožiny *A* do každého vrcholu patriaceho do podmnožiny *B*. Z toho dôvodu existuje $m \cdot n$ hrán.

Na začiatku je pridelený všetkým hodnotiteľom spolu *K* bodov. Tieto sa rovnomerne rozdelia medzi m hodnotiteľov. Teda každý z hodnotiteľov dostane rovnaký počet bodov, a to konkrétne $\frac{K}{m}$. Uvedený počet bodov označíme ako počet v nultej iterácii, keďže uvedený výpočet sa vykonáva iba raz na začiatku. Počet bodov, ktoré majú k dispozícii hodnotitelia alebo hodnotení, budeme zapisovať ako a_{ij}, respektíve b_{ij}, kde *i* predstavuje poradové číslo hodnoteného alebo hodnotiteľa a *j* predstavuje poradové číslo iterácie, ktorá práve prebieha. Písmeno *a* alebo *b* je pritom pridelené na základe toho, do ktorej skupiny uvedený subjekt patrí, teda či sa jedná o hodnotiteľa alebo hodnoteného.

Ďalším dôležitým aspektom tohto modelu je definovanie váh. Každá hrana z množiny hrán E má pritom dve váhy, a to z toho dôvodu, že v celkovom pohľade žiadna z hrán nie je

smerovo orientovaná a teda je možné po nej presúvať body (hodnotenie) tak od hodnotiteľa ako aj hodnoteného. Avšak pri výpočte modelu uvažujeme vždy iba o jednom smere, a teda celkový graf je rozčlenený na 2 bipartitné grafy, ktorých hrany sú smerovo orientované. Na základe toho potom určujeme váhy jednotlivých hrán od každého subjektu. Váhy na jednotlivých hranách označíme ako wa_{uv}, kde u predstavuje poradové číslo hodnotiteľa a v predstavuje poradové číslo hodnoteného. Tak teda dostaneme napríklad wa_{13}, čo znamená, že predstavuje váhu pre hranu, po ktorej hodnotiteľ 1 odovzdáva v modeli zvolený počet bodov hodnotenému 3. Obdobne pre hodnotených je definovaná váha na každej hrane ako wb_{vu}, kde v predstavuje poradové číslo hodnoteného a u predstavuje poradové číslo hodnotiteľa. Teda v danom modeli existuje množina váh Wa a množina váh Wb, ktoré pozostávajú z $m \cdot n$ prvkov a obe množiny sú rovnako mohutné.

Váhy pre všetky subjekty z rovnakej podmnožiny sú pritom pevne definované pre celú dobu iterovania. Priraďovanie váh sa realizuje prostredníctvom určovania preferencií, kde každý z hodnotiteľov urči svoje preferencie voči hodnoteným, to znamená, že ich zoradí v poradí od najlepšieho k najhoršiemu, respektíve mu priradí poradové číslo. Na základe tohto poradového čísla bude každá z hrán, ktorá z vrcholu tohto hodnotiteľa vychádza, zaťažená váhou, ktorá prislúcha danému poradiu zvoleného hodnoteného. Uvedené váhy pritom budú vopred pevne dané a nebudú sa meniť. Postup výpočtu váh ako aj ich stanovovanie bude predmetom samostatnej podkapitoly, preto tu tento postup definujeme iba teoreticky a bližšie mu nie je venovaná pozornosť.

Na základe uvedených definícií, ktorými sme stanovili jednotlivé premenné v našom modeli, môžeme stanoviť presný postup výpočtu počtu bodov, ktoré získajú jednotliví hodnotení, respektíve hodnotitelia. Uvedený postup má iteračný charakter, pričom počet opakovaní nie je určené vopred. Tento počet iterácií môže byť definovaný samotným kontrolórom modelu v podmienkach konkrétneho podniku alebo prostredníctvom limitného kritériá, ktoré automaticky ukončí uvedený program.

Postup bodovacieho algoritmu môžeme teda zapísať nasledovne. V nultej iterácii dôjde k výpočtu bodov, ktoré získajú hodnotitelia na začiatku.

$$a_{i0} = \frac{K}{m}, \text{ pre } i = 1,\ldots,m \tag{4}$$

V prvej iterácii dôjde k prerozdeleniu bodov od hodnotiteľov k hodnoteným. K tomuto prerozdeleniu teda použijeme váhy, ktoré sme predtým definovali. Teraz je potrebné pre každého hodnoteného vypočítať, koľko bodov získa v prvej iterácii. To zapíšeme ako:

$$b_{i1} = \sum_{s=1}^{m} \left(a_{s0} \cdot wa_{si}\right), \text{ pre } i = 1,\ldots,n \tag{5}$$

V ďalšom pokračovaní prvej iterácie musíme vypočítať počet bodov, ktoré budú spätne prerozdelené hodnotiteľom. To uskutočníme obdobne. V tomto prípade však budeme používať váhy z podmnožiny *Wb*. A teda dostaneme:

$$a_{i1} = \sum_{t=1}^{n} \left(b_{t1} \cdot wb_{ti}\right), \text{ pre } i = 1,\ldots,m \tag{6}$$

Teraz nasleduje druhá iterácia a teda opätovný výpočet počtu bodov, ktoré získa každý z hodnotených. To zobrazíme obdobne ako v prvej iterácii.

$$b_{ip} = \sum_{s=1}^{m} \left(a_{s\,p-1} \cdot wa_{si}\right), \text{ pre } i = 1,\ldots,n \tag{7}$$

A teda v ľubovoľnej *p*-tej iterácii je počet bodov hodnotiteľa rovný:

$$a_{ip} = \sum_{t=1}^{n} \left(b_{tp} \cdot wb_{ti} \right), \text{ pre } i = 1, \ldots, m \tag{8}$$

Na základe vzťahov (11) a (13) si definujeme celý algoritmus jednej iterácie, ktorý určí počet bodov pre hodnotiteľa a hodnoteného na konci iterácie. To uvádzame v nasledujúcich dvoch vzorcoch (14) a (15).

$$a_{ip} = \sum_{t=1}^{n} \left(wb_{ti} \cdot \sum_{s=1}^{m} \left(a_{i\,p-1} \cdot wa_{is} \right) \right), \text{ pre } i = 1, \ldots, m \tag{9}$$

$$b_{ip} = \sum_{s=1}^{m} \left(wa_{si} \cdot \sum_{t=1}^{n} \left(b_{tp-1} \cdot wb_{ti} \right) \right), \text{ pre } i = 1, \ldots, n \tag{10}$$

Pre prípad výpočtu alternatívnej premennej prostredníctvom jednej iterácie je nasledovné:

$$a_{ip} = \sum_{t=1}^{n} \left(wb_{ti} \cdot \sum_{s=1}^{m} \left(wa_{is} \cdot \sum_{u=1}^{n} \left(b_{up-1} \cdot wb_{ui} \right) \right) \right), \text{ pre } i = 1, \ldots, m \tag{11}$$

$$b_{ip} = \sum_{s=1}^{m} \left(wa_{si} \cdot \sum_{t=1}^{n} \left(wb_{ti} \cdot \sum_{u=1}^{m} \left(a_{u\,p-2} \cdot wa_{ui} \right) \right) \right), \text{ pre } i = 1, \ldots, n \tag{12}$$

Uvedené váhy je možné definovať ako zobrazenie množín *Va* a *Vb*. Uvedené množiny sa vytvárajú cez sériu 4 pravidiel:

1. Súčet váh musí byť rovný jednej, teda:

$$\sum_{i=1}^{n} va_i = 1 \quad \wedge \quad \sum_{i=1}^{m} vb_i = 1 \tag{13}$$

2. Prvky množiny váh *Va* aj *Vb* musia vytvárať nerastúcu postupnosť a teda musí platiť:

$$va_i - va_{i+1} \geq 0 \quad pre\, i = \{1, \ldots, n-1\} \tag{14}$$

$$vb_i - vb_{i+1} \geq 0 \quad pre\, i = \{1, \ldots, m-1\} \tag{15}$$

3. Pre všetky základné váhy z množín *Va* a *Vb* platí:

$$va_i \geq 0 \quad pre\, \forall i; i = \{1, \ldots, n\} \tag{16}$$

$$vb_i \geq 0 \quad pre\, \forall i; i = \{1, \ldots, m\} \tag{17}$$

4. Ak platia podmienky 1 a 3, potom nutne musí platiť, že akýkoľvek prvok množiny *Va* alebo *Vb* má číselnú hodnotu menšiu alebo rovnú ako 1.

$$va_i \leq 1 \quad pre\, \forall i; i = \{1, \ldots, n\} \tag{18}$$

$$vb_i \leq 1 \quad pre\, \forall i; i = \{1, \ldots, m\} \tag{19}$$

Veta 1

Ak platí, že súčet postupnosti je rovný 1 a zároveň platí, že každý z členov postupnosti je nezáporný, potom platí, že každý z členov postupnosti je menší, nanajvýš rovný 1.

Dôkaz

Predpokladajme, že sú splnené podmienky1 a 3 a zároveň neplatí podmienka 4. Teda existuje taká množina $Va = \{va_1, va_2, \ldots, va_n\}$, že:

1. $\sum_{i=1}^{n} va_i = 1$
2. $va_i \geq 0 \quad \forall i \; i = \{1,2,\ldots,n\}$

Ale existuje také $j \in \{1,2,\ldots,n\}$ také, že $va_j > 1$.

Keďže operácia súčtu reálnych čísel je komutatívna, to znamená, že celkový výsledok súčtu nie je závislý na poradí sčítancov, môžeme bez ujmy na všeobecnosti (BUNV) preusporiadať členy postupnosti tak, aby va_j bol prvý člen postupnosti. Ostatné členy sú potom v rovnakom poradí.

$$\sum_{i=1}^{n} va_i = va_j + \sum_{i=2}^{n} va_i \tag{13}$$

Platí, že $\sum_{i=2}^{n} va_i \geq 0$, pretože platí podmienka 2, ktorá tvrdí, že $va_i \geq 0 \quad \forall i \; i = \{1,2,\ldots,n\}$. Ďalej vieme, že $va_j > 1$. A preto súčet kladného čísla a čísla väčšieho ako 1 je určite číslo ostro väčšie ako 1, teda

$$\sum_{i=1}^{n} va_i > 1, \tag{14}$$

čo je v spore s podmienkou 1, že suma prvkov postupnosti má byť rovná jednej.

q.e.d.

Vzhľadom na náročnosť výpočtov z hľadiska množstva výpočtových úkonov je použitie uvedeného modelu bez využitia výpočtovej techniky veľmi obmedzené. Na uvedený model z toho dôvodu vznikol pilotný program, ktorý je schopný vypočítať výsledky zo zvolených vstupných hodnôt, čo zjednodušuje použitie modelu a tým pádom je tento model možné využiť v podnikovej praxi.

4. Záver

V tomto článku sme sa zaoberali návrhom modelu hodnotenia spokojnosti zamestnancov a jeho použitím cez výpočtovú techniku. Uvedený model hodnotenia je síce objemom výpočtov náročný, avšak použitie výpočtovej techniky umožní jeho využitie aj v podnikovej sfére. Prínos tohto modelu je práve v oblasti spresnenia hodnotenia spokojnosti. Hodnotenie spokojnosti zamestnancov je pritom dôležitá činnosť za účelom spravodlivého ohodnotenia zamestnancov.

5. Literatúra

[1] CARTER, L. – ULRICH, D. – GOLDSMITH, M.: *Best Practices in Leadership Development and organization Change*. New Jersey : John Wiley and Sons. 2004. ISBN 0-7879-7625-3.

[2] FARNHAM, D.: *Employee Relations in Context*. London: Institute of Personnel and Development. 2000.

[3] FUCHSOVÁ, K. - KRAVČÁKOVÁ, G.: *Manažment pracovnej motivácie*. Bratislava: Iris, 2004.

[4] HIRSCHMAN, Albert O.: The Paternity of an Index. In: *The American Economic Review*. American Economic Association. Vol. 54. 1964. No. 5. p. 761
[5] KOLLÁRIK, T.: *Sociálna psychológia práce.* Bratislava: VYDAVATEĽSTVO UK, 2002.
[6] MATZLER, K. – RENZL, B.: Assessing asymmetric Effects in the Formation of Employee Satisfaction. In: *Tourism Management*. Vol. 28. 2007. s. 1093-1103.

Adresa autorov:
Zuzana Hajduová, RNDr. PhD.
Ekonomická univerzita v Bratislave
Podnikovohospodárska fakulta v Košiciach
Tajovského 13, 040 01 Košice
zuzana.hajduova@euke.sk

Marek Andrejkovič, Ing. PhD.
Ekonomická univerzita v Bratislave
Podnikovohospodárska fakulta v Košiciach
Tajovského 13, 040 01 Košice
marek.andrejkovic@gmail.com

Tento príspevok je výstupom projektu VEGA 1/0339/10 – „Ekonomický rast a jeho limitujúce faktory - Návrh nových ekonomických cieľov a indikátorov kvality života potrebných pre vytvorenie všeobecne platnej, novej metodiky hodnotenia kvality života a trvalo udržateľného rozvoja na území Slovenskej republiky“.

Expoziční křivky a neparametrická regrese
Exposure curves and nonparametric regression

Jan Hrevuš

Abstrakt: Expoziční křivky hrají důležitou roli při výpočtu cen zajištění škodního nadměrku. Historicky první výzkum na toto téma byl proveden v roce 1963, v němž Ruth E. Salzmannová popsala vztah mezi akumulovanou výší pojištěných škod a procentní výší pojistné částky. Diskrétní empirické rozdělení škod bylo odvozeno na základě historických škod americké pojišťovny „Insurance Company of North America“ vzniklých v roce 1960. I přes poměrně velký rozsah základního souboru je třeba data vyhladit pomocí regrese, aby následný výpočet ceny neproporčního zajištění pomocí tabelovaných hodnot byl více legitimní. Jednou z možností je užít neparametrickou regresi, nicméně tato metoda obvykle nebývá obsažena ve statistických paketech a musí být programována manuálně. Jednou z výjimek, kde je metoda automaticky obsažena, je statistický software XLSTAT 2010, pomocí něhož budou provedeny všechny v tomto článku uvedené výpočty.

Abstract: Exposure curves play a significant role in pricing of property excess-of-loss reinsurance. The first research on that topic was introduced by Ruth E. Salzmann in 1963 in her paper called “Rating by Layer of Insurance” where an accumulated loss cost distribution by percentage of insured value was developed. An empirical discrete loss cost distribution was created based on the direct loss data of the Insurance Company of North America for 1960 incurred year. Although the sample of the data was relatively large, some sort of regression to smooth the observed values seems to be appropriate. One of the possible directions might be use of nonparametric regression, however, that method is not usually included in almost any statistical software and usually has to be programmed manually. All calculations in this paper will be done in software XLSTAT 2010 which includes that procedure.

Klíčová slova: neparametrická regrese, expoziční křivky, zajištění, Salzmann.

Key words: Nonparametric regression, Exposure curves, Reinsurance, Salzmann.

JEL classification: C14, G22

1. Introduction

Nonparametric regression is still a neglected discipline in respect of both practical use and tuition at many universities. The logical consequence of that is a lack of this method in an absolute majority of statistical software packages. One of the few exceptions is statistical software called XLSTAT. It is an user friendly add-in for MS Excel which includes many statistical methods including nonparametric regression. All calculations in this paper will be done in XLSTAT 2010.

The method analyzed concerns to the reinsurance area, where property non-proportional reinsurance treaties might by priced by various exposure curves. That approach was developed in 1963 and will be further discussed and described.

2. Basic ideas of nonparametric regression

The nonparametric regression can be used as an alternative to parametric approach in case of violation of the principal assumptions about the use of the parametric regression. Further reasons for using of the nonparametric approach might be detection of outliers or

making a prior assumption about the shape of the regression curve for further application of the parametric approach.

There are more different types of nonparametric regression, however, only the type called *kernel regression* will be described in this article. Further, only the simple model with one quantitative response variable Y and one quantitative predictor variable X will be discussed, however, all further used formulas can be easily modified for multiple nonparametric regression. The simple regression function described in this article has the following form:

$$Y = f(X) + \varepsilon. \tag{2}$$

Kernel regression is a form of locally weighted averaging and belongs to the family of smoothing techniques. The basic idea of that method is computing estimator $\hat{y}_0$ as a weighted average from the nearby lying observations. Assuming there is a focal value x_i for each observation y_i then the closer the focal value x_i of the observation y_i to the focal value x_0 the higher impact of y_i on $\hat{y}_0$.

The essential issue with a huge impact on final result is a suitable choice of the number of observations used for the local averaging where more possible methods exist.

One of the options might be use of fixed number of the nearest observations. This method is called *nearest neighbor estimate.*

Another option might be a fixed distance from the focal value x_0 where this distance is called *bandwidth* of the kernel estimator and will be further marked as *h*.

The other characteristic of kernel regression is the use of a *kernel function* to weight the observations of the learning sample depending on their distance from the focal value x_0 within the bandwidth *h* for prediction of $\hat{y}_0$. All kernel functions are constructed in the same way where the closer the value x_i to the value x_0 the higher impact of y_i on prediction of $\hat{y}_0$. Many kernel functions have been suggested (e.g. Uniform, Triangle, Epanechnikov, Quartic, Triweight, Tricube, Gaussian, Cosine, etc.), however, it has been proved that choice of the kernel function is not so important as all of them provide very similar results.

Two of the popular kernel functions are the *Epanechnikov* and the *tricube kernels* with the following functions:

$$K(z) = \begin{cases} 0{,}75(1 - z^2) & for\ |z| \leq 1 \\ 0 & for\ |z| > 1 \end{cases} \tag{2}$$

for the *Epanechnikov* kernel and

$$K(z) = \begin{cases} (1 - |z|^3)^3 & for\ |z| < 1 \\ 0 & for\ |z| \geq 1 \end{cases} \tag{3}$$

for the *tricube kernel*, where

$$z_i = \frac{(x_i - x_0)}{h}. \tag{4}$$

Both kernel functions are shown on figure 1.

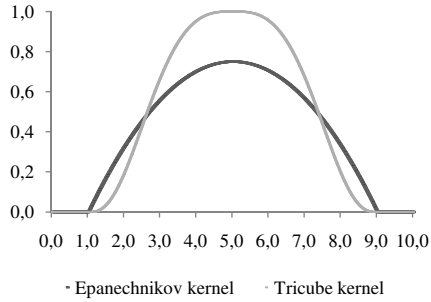


Fig. 1. Epanechnikov kernel (black line) and tricube kernel (grey line) for bandwidth $\boldsymbol{h = 4}$ ***and*** $\boldsymbol{x_i = 5}$ **.**

As the kernel function has to be calculated for each observation separately the kernel regression is very demanding for computation.

Having chosen the kernel function, the weight of each observation y_i further used for estimation of $\hat{y}_0$ is calculated as

$$w_{0i} = \frac{K(z_{0i})}{h} \ , \tag{5}$$

where

$$z_{0i} = \frac{(x_i - x_0)}{h}. \tag{6}$$

and then having the weights w_{0i} for each observation x_i , where $i = 1, 2, \dots, n$, it is already possible to calculate the estimation of the response variable $\hat{y}_0$ corresponding to the value of the predictor variable x_0 using the *Nadaraya-Watson* formula:

$$\hat{y}_0 = \hat{f}(x_0) = \frac{\sum_{i=1}^{n} w_{0i} y_i}{\sum_{i=1}^{n} w_{0i}} \ . \tag{7}$$

For any analysis made by kernel regression, it is the most important step to select an appropriate bandwidth. The aim of the nonparametric regression is to find the smallest value of *h* which provides a smooth fit. Unfortunately, there is a contradiction between these two targets as the larger the bandwidth, the smoother the regression curve. On the other hand the narrower the bandwidth, the closer the estimators to the observed values.

Experience has shown that probably the most effective decision criterion is to compare a set of visual trials produced by computers. Visual comparison of the fitted values to the observed values remains due to its simplicity the most frequently used method. The obvious disadvantages are the subjectivity and impossible use for multivariable models.

Another approach might be the method called *cross-validation*. The key idea is to skip the *i-th* observation at the focal value x_i when calculating the local regression. Assuming $\hat{y}_{i(-i)}$ is the fitted value whose calculation is based on $n - 1$ observations and omitting the *i-th* observation, the cross-validation function is defined as:

$$\mathrm{CV(h)} = \frac{\sum_{i=1}^{n}(\hat{y}_{i(-i)} - y_i)^2}{n} \quad . \tag{7}$$

The target is to find the value h which minimizes $\mathrm{CV(h)}$. In practice it is necessary to compute $\mathrm{CV(h)}$ for a wide range of values of h and choose the most appropriate one. It is important to keep in mind that $\mathrm{CV(h)}$ is only an estimate and therefore a subject to sampling variation, particularly in small samples.

3. Salzmann exposure curves

The first idea about exposure curves was introduced by Ruth E. Salzmann in her revolutionary article called "Rating by Layer of Insurance" [9] from 1963 and further discussed by Stephen J. Ludwig [9].

Salzmann based her study on the homeowners fire losses portfolio of the Insurance Company of North America for 1960 incurred year as of May 31, 1961. She developed various discrete accumulate loss cost distributions by percentage of insured value for several types of buildings. The distributions were based on aggregated observed values for four types of buildings: frame – protected, frame – unprotected, brick – protected and brick – unprotected.

Based on the discrete data [9] and using the kernel regression, it was possible to develop following exposure curves for several types of buildings.

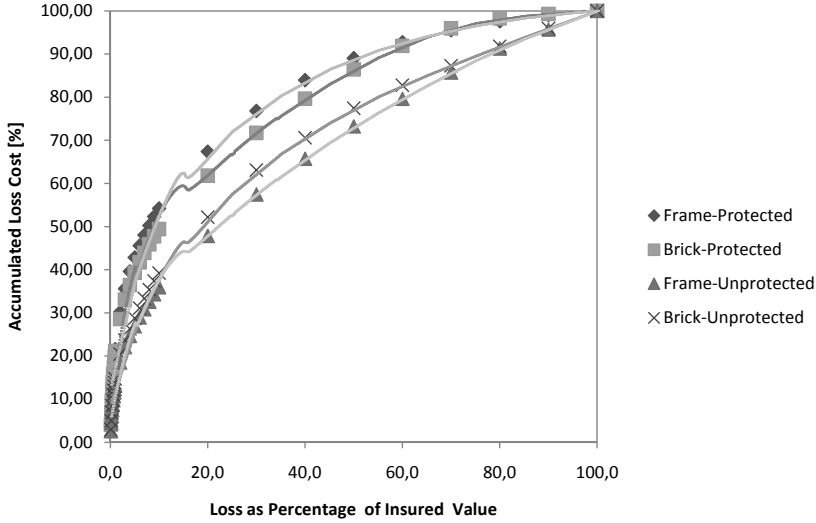


Fig. 2. Exposure curves calculated by kernel regression (h = 15; Epanechnikov kernel) based on Salzmann´s data [9]

The exposure curves are often used for *exposure pricing* of property non-proportional reinsurance treaties. Basically, the closer the curve to the diagonal the more total losses the portfolio contains and on the other hand the more far the curve from the diagonal the more partial losses the portfolio contains. There are many exposure curves used in praxis, probably the most frequently used are called *"Swiss Re exposure curves"*[5]. The curves close to the

diagonal are used mainly for personal lines where the total losses are expected while the curves more far away from the diagonal are used for industrial lines of business with higher sums insured where the total loss is not expected due to the various fire protections and size of the risk.

4. Conclusion

The exposure curves for personal lines were developed based on Salzmann´s data using the nonparametric regression. The article aimed to prove that nonparametric regression is an suitable alternative approach to the parametric approach, especially to the nonlinear regression which is usually used for this purpose and that the nonparametric approach can be used for further development of exposure curves based on more recent data. As discussed in the article, the nonparametric regression is a very subjective method and the final result depends highly on the final opinion of the user. The only problematic part when estimating the curves is the part close to the observations with the focal value $x_i = 14$ where all the curves have an unexpected shape. Those parts of the curves have no logical explanations and it is caused by not uniformly distributed observations and the sensitivity of the selected method on that. The problem would be solved by having the complete database of losses what would result to more observations for fitting the curve.

5. Literature

[1]Addinsoft, http://www.xlstat.com/

[2]Bernegger, S.: The Swiss Re exposure curves and MBBEFD distribution class. In: AUSTIN Bulletin, vol. 27, 1997, p. 99 - 111

[3]Cipra, T.: Zajištění a přenos rizik v pojišťovnictví. Grada Publishing, Praha (2004)

[4]Fox, J.: Nonparametric Simple Regression: Smoothing Scatterplots. Sage Publications, Inc., Thousand Oaks (2000)

[5]Guggisberg, D.: Exposure rating. Swiss Re, Zurich (2004)

[6]Härdle, W.: Applied nonparametric regression. Cambridge University Press, Cambridge (1995)

[7]Hebák, P., Hustopecký, J., Malá, I.: Vícerozměrné statistické metody (2). Informatorium, Praha (2005)

[8]Hrevuš, J.: Economic performance of different countries and life expectancy of their inhabitants. In: Applications of Mathematics and Statistics in Economy, 2010

[9]Ludwig, J.S.: An Exposure Rating Approach to Pricing Property Excess-of-Loss Reinsurance. In: Proceedings of the Casualty Actuarial Society, vol. LXXVIII, 1991, p. 110 - 145

[10] Salzmann, E. R.: Rating by Layer of Insurance. In: Proceedings of the Casualty Actuarial Society, vol. L, 1963, p. 15 – 26.

Adresa autora:

Jan Hrevuš, Ing.
VIG Re zajišťovna, a.s.
Klimentská 46, 110 00 Praha 1
j.hrevus@vig-re.com

Indikátory Evropa 2020 a možnosti redukce proměnných
Indicators Europe 2020 and the Methods of Data Reduction

Lenka Hudrlíková

Abstract: Europe 2020 is the EU‘s strategy for smart, sustainable and inclusive growth for the coming decade. Assessment of these targets is monitored on the basis of the set of indicators called Europe 2020. Cross-country comparison is possible only by a multilateral approach. The way to deal with this multidimensional view is a constructing of a single composite indicator. It integrates information from all sub-indicators. Data reduction methods, used in the paper, are the principal component analysis and factor analysis. These methods are used for the assessment of the structure of relations indicators Europe 2020 and the possibility of their aggregation into a single composite indicator.

Key words: Indicators Europe 2020, Principal Component Analysis, Factorial analysis, Composite indicator

Klíčová slova: Ukazatelé Evropa 2020, Metoda hlavních komponent, Faktorová analýza, kompozitní indikátor

JEL classification: C38, O11

1. Úvod

Evropská komise v roce 2010 přijala novou strategii nazvanou Strategie pro inteligentní a udržitelný růst podporující začlenění. Strategie je orientována na několik hlavních cílů, jež by měly ekonomiky členských zemí dosáhnout. Jedná se o pět hlavních cílů:

- Zvýšit zaměstnanost u osob ve věkové skupině 20 až 64 let na nejméně 75 %,
- Investovat 3 % HDP do výzkumu a vývoje,
- Snížit emise skleníkových plynů oproti úrovním roku 1990 nejméně o 20 %, zvýšit podíl obnovitelných zdrojů energie v konečné spotřebě energie na 20 % a zvýšit energetickou účinnost také o 20 %,
- Snížit podíl dětí, které předčasně ukončují školní docházku, na 10 % a zvýšit podíl obyvatel ve věkové skupině od 30 do 34 let, kteří mají ukončené terciární vzdělání, na 40 %,
- Snížit o 25 % počet obyvatel, kteří žijí pod hranicí chudoby.

Zároveň byla vytvořena sestava ukazatelů, jež mají sloužit k dohlížení na plnění cílů. Hlavních ukazatelů je osm a jsou rozděleny do tří oblastí – ekonomické, sociální a environmentální. Tato sestava ukazatelů je ještě rozšířena o další podpůrné ukazatele. Z hlediska mezinárodního srovnání zemí vzniká problém vícestranného srovnávání pomocí několika ukazatelů. Jedním z přístupů, jak se vypořádat s tímto multidimezionálním pohledem, je vytvoření jednoho kompozitního indikátoru, který v sobě bude zahrnovat informace ze všech dílčích ukazatelů. Pomocí vícerozměrných statistických metod lze posoudit, zda jsou ukazatele Evropa 2020 vhodné k agregaci a k vytvoření jednoho kompozitního indikátoru.

2. Data a metodika

Sestavu osmi hlavních ukazatelů Evropa 2020 tvoří ukazatele dle [3]:

- Míra zaměstnanosti osob ve věku 20-64 let (EMP)
- Hrubé domácí výdaje na výzkum a vývoj jako podíl HDP (GERD)
- Emise skleníkových plynů (základ rok 1999) (GH)

- Podíl obnovitelných zdrojů na hrubé konečné spotřebě energie (RE)
- Energická náročnost ekonomiky (EN)
- Studenti ve věku 18-24 předčasně ukončující docházku (EL)
- Podíl osob v populaci ve věku 30-34, kteří úspěšně ukončili terciární vzdělání (TE)
- Podíl osob v populaci žijících na hranici chudoby a sociálního vyloučení (POV)

Tyto ukazatele jsou harmonizovány pro všechny členské státy EU. Lze je používat v mezinárodním srovnávání zemí. Data použitá pro výpočty jsou z databáze EUROSTATu. Za všechny členské země jsou poslední dostupná data za rok 2008. Pro účel mezinárodního srovnávání zemí podle ukazatelů Evropa 2020 je použita metoda hlavních komponent a faktorová analýza. Model faktorové analýzy lze chápat jako zobecnění modelu komponentní analýzy. Výpočty byly pořízeny v programu Statistica 9. Obě analýzy v některých případech mohou vést ke značné redukci proměnných v úloze. Ojediněle by se mohlo jednat i o redukci na pouze jednu hlavní komponentu, resp. faktor. Pokud data nejsou vhodná pro tuto redukci, využití těchto statistických metod spočívá v průzkumové analýze dat před samotnou konstrukcí kompozitního indikátoru. A v neposlední řadě při určení vah, jež budou mít dílčí ukazatelé při agregaci do jednoho kompozitního ukazatele.

3. Analýza a výsledky

Analýza byla provedena ve třech hlavních krocích:

1. Spočítání korelační matice, pokud jsou korelace mezi ukazateli malé, je nepravděpodobné, že bude nalezen společný faktor.
2. Identifikace vhodného počtu komponent (faktorů).
3. Rotace faktorů pro posílení možnosti interpretace výsledků.

Pro ověření předpokladů pro užití metody hlavních komponent a faktorové analýzy je třeba ověřit, zda proměnné jsou či nejsou vzájemné korelované.

Tabulka 1: Korelační matice

	EMP	**GERD**	**GH**	**RE**	**EN**	**EL**	**TE**	**POV**
EMP	1,000	-0,073	-0,095	-0,457	0,241	-0,037	-0,142	-0,110
GERD	-0,073	1,000	0,018	0,334	-0,115	0,207	-0,058	0,350
GH	-0,095	0,018	1,000	-0,071	0,215	0,240	-0,115	-0,089
RE	-0,457	0,334	-0,071	1,000	-0,373	-0,012	0,184	0,221
EN	0,241	-0,115	0,215	-0,373	1,000	0,190	-0,495	-0,252
EL	-0,037	0,207	0,240	-0,012	0,190	1,000	-0,273	-0,044
TE	-0,142	-0,058	-0,115	0,184	-0,495	-0,273	1,000	0,196
POV	-0,110	0,350	-0,089	0,221	-0,252	-0,044	0,196	1,000

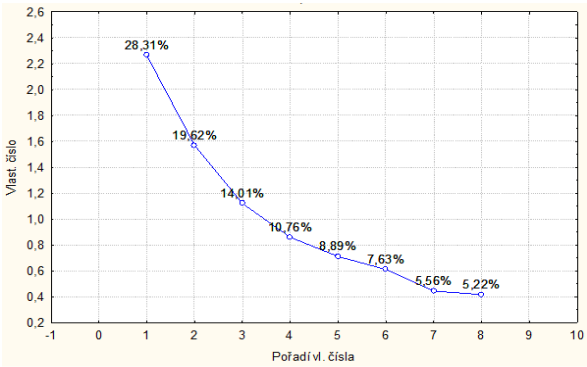
Dle tabulky 1 žádný z korelačních koeficientů nenaznačuje silnou závislost mezi ukazateli. Největší záporná korelace (tj. - 0,495) je mezi podílem dětí, kteří předčasně ukončují školní docházku (TE) a energetickou náročností ekonomiky (EN). Slabší nepřímá závislost -0,457 je mezi mírou zaměstnanosti (EMP) a podílem obnovitelných zdrojů (RE). Jelikož jsou hodnoty obou korelačních koeficientů blízké 0,5, korelaci lze považovat za slabou. Navíc v obou případech chybí racionální ekonomické zdůvodnění závislosti. Použití zmíněných vícerozměrných statistických metod tedy nemusí přímo přinášet žádné využitelné výsledky, jelikož nemusí dojít k dostatečné redukci proměnných. Protože různé ukazatele vždy reprezentují určitou oblast (ekonomická, sociální, environmentální), lze i přes slabou korelaci předpokládat, že mohou existovat skryté (latentní) veličiny, tzv. hlavní komponenty nebo faktory, což jsou lineární kombinace původních ukazatelů. Data nejsou ve stejných měřicích jednotkách; pro analýzu je tedy využita korelační matice (hodnoty korelačních koeficientů

v tabulce 1), resp. kovariační matice standardizovaných proměnných s nulovou střední hodnotou a jednotkovým rozptylem. Nové proměnné, tzv. komponenty, by měly vysvětlovat variabilitu původních proměnných. U faktorové analýzy by vytvořené faktory měly co nejlépe reprodukovat vzájemné lineární vztahy původních proměnných. Pro stanovení hlavních komponent resp. faktorů analýza využívá vlastní čísla. Vlastní čísla v tabulce 2 jsou získaná z korelační matice pro 8 proměnných (korelační koeficienty ukazatelů Evropa 2020) v tabulce 1. Součet hodnot vlastních čísel se rovná počtu proměnných.

Tabulka 2: Vlastní čísla, % celkové rozptylu

pořadí vl. čísla	vlastní číslo	% celkového rozptylu	kumulativní vlastní číslo	kumulativní % celk. rozptylu
1	**2,265**	28,308	2,265	28,308
2	**1,570**	19,619	3,834	47,926
3	**1,120**	14,006	4,955	61,932
4	0,861	10,758	5,815	72,690
5	0,712	8,894	6,527	81,584
6	0,611	7,634	7,137	89,218
7	0,445	5,558	7,582	94,777
8	0,418	5,223	8,000	100,000

Kritérií pro rozhodnutí o vhodném počtu nových proměnných je několik a jsou vesměs založeny na vlastních číslech. Nejčastější doporučení jsou dle [1]: (i) vynechat všechny faktory s vlastním číslem menším než 1, (ii) faktory by měly vysvětlovat 80-90 % rozptylu, (iii) vynechat všechny faktory s vlastními čísly pod 0,7, (iv) počet faktorů by měl být dostatečně malý, aby výsledky byly srozumitelné, (v) pomocí grafu. První hlavní komponenta vysvětluje pouze 28,3 % z celkového rozptylu. Druhá komponenta vysvětluje maximum ze zbývajícího rozptylu, tj. 19,62 % z celkového rozptylu. Třetí a čtvrtá komponenta mají vlastní číslo blízké 1. Pro rozhodnutí o počtu nových proměnných lze také podle grafu tzv. scree plot na obrázku 1. Graf znázorňuje vlastní čísla, která udávají, kolik z celkové variability je vysvětleno pomocí faktoru.



Obrázek 1: Vlastní čísla

Výsledky v tabulce 2 stejně jako v grafu 1 indikují, že počet hlavních komponent by měl být 3 - 4. Tři faktory vysvětlují však pouze 62 % celkové variability. Čtyři faktory však už bývají nesrozumitelné při interpretaci výsledků. Určení počtu komponent není jednoznačné.

Korelační koeficienty mezi hlavními komponenty (faktory) a původními proměnnými jsou komponentní (faktorové) zátěže uvedené v tabulce 3. Kvadrát komponentní zátěže vyjadřuje, kolik procent z rozptylu je vysvětleno hlavní komponentou.

Tabulka 3: Komponentní (faktorové) zátěže

	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5	Faktor 6	Faktor 7	Faktor 8
EMP	**-0,510**	-0,306	**-0,607**	-0,209	-0,203	-0,311	-0,064	0,303
GERD	0,363	**0,624**	-0,480	0,006	0,003	-0,369	0,124	-0,310
GH	-0,266	0,463	0,464	**-0,623**	0,186	-0,237	-0,135	0,060
RE	**0,705**	0,332	0,184	0,331	0,067	-0,226	0,062	0,436
EN	**-0,765**	0,221	-0,053	0,083	0,341	0,083	0,473	0,097
EL	-0,247	**0,698**	0,020	-0,034	-0,598	0,285	0,047	0,090
TE	**0,604**	-0,431	0,153	-0,432	-0,255	-0,040	0,416	-0,005
POV	**0,542**	0,199	-0,495	-0,354	0,304	0,425	-0,065	0,138

Hodnoty v tabulce 3 v absolutní hodnotě vyšší než 0,05 vyjadřují korelaci mezi původním ukazatelem a hlavní komponentou (faktorem). Kromě ukazatele emise skleníkových plynů (GH) jsou zátěže u všech ostatních indikátorů nejvíce korelovány s některým z prvních třech faktorů. To opět značí, že správný počet komponent (faktorů) by měl být tři nebo čtyři. Čím více hlavních komponent je potřeba, tím méně je však metoda užitečná. Pro zodpovězení otázky, kolik hlavních komponent je třeba uvažovat bez větší ztráty informace, je lepší provést také faktorovou analýzu s rotací faktorů. Při rotaci faktorů se z dané matice faktorových zátěží (tabulka 3) získá nová matice. V tabulce 4 jsou výpočty při použití metody Varimax. Tučně označené jsou zátěže větší než 0,5.

Tabulka 4: Faktorové zátěže pro tři faktory, extrakce: hlavní komponenty, rotace Varimax

	Faktor 1	Faktor 2	Faktor 3
EMP	0,003	0,007	**-0,850**
GERD	**0,845**	0,167	0,104
GH	-0,194	**0,587**	0,343
RE	0,396	-0,175	**0,674**
EN	-0,214	**0,649**	-0,412
EL	0,247	**0,688**	0,124
TE	-0,048	**-0,689**	0,310
POV	**0,710**	-0,270	0,043
Výkl.roz	1,521	1,845	1,588
Prp.celk	0,190	0,231	0,198
Kum. Prp.celk	0,190	0,421	0,619

První faktor má vysokou faktorovou zátěž u proměnných výdaje na výzkum a rozvoj (GERD) a podíl populace žijící na hranici chudoby (POV). Faktoru 2 dominují ukazatelé energetická náročnost (EN), osoby předčasně ukončující školní docházku (EL), podíl populace s úspěšně ukončeným terciárním vzděláním (TE) a emise skleníkových plynů (GH). Existují i jiné možnosti rotace faktorů. V tomto případě však nepřináší příliš jiné výsledky, čili závěry o použití analýzy jsou podobné.

Aplikování metody hlavních komponent a faktorové metody na data souboru ukazatelů Evropa 2020 nevede k určení správného počtu nových umělých proměnných. Nové proměnné u komponentní analýzy mají co nejvíce vyčerpat maximum celkového rozptylu původních proměnných, resp. maximálně reprodukovat celkovou korelační matici původních proměnných. U faktorové analýzy mají nové proměnné vysvětlovat závislosti mezi proměnnými. Z uvedených výsledků analýzy plyne, že nelze splnit tyto cíle a zároveň dostatečně redukovat počet proměnných.

4. Závěr

Ukazatele Evropa 2020 nejsou vhodné pro redukci proměnných, tzn. počtu ukazatelů. Metody se neprojevily jako užitečné pro redukci uvedených dat, což je způsobeno hlavně nekorelovaností jednotlivých ukazatelů. Tato vlastnost se jeví jako nevýhodná, jejím důsledkem totiž je, že pomocí užitých metod redukce dat nelze proměnné (ukazatelé Evropa 2020) vyjádřit pomocí latentních neměřitelných proměnných (přestože ukazatele zastupují různé oblasti). Tato vlastnost je však naopak vítána při samotné konstrukci kompozitního indikátoru. Uvedené analýzy jsou cenné pro posouzení struktury ukazatelů již před samotnou konstrukcí kompozitního indikátoru. Při konstrukci kompozitních ukazatelů je lépe využívat metodu hlavních komponent než faktorovou analýzu, a to vzhledem k její jednoduchosti. Po posouzení vhodnosti dílčích ukazatelů a jejich vlivu tato analýza pomáhá konstruovat váhy reprezentující informace obsažené v dílčích ukazatelích.

5. Literatura

[1] HEBÁK, PETR A KOL. Vícerozměrné statistické metody (3). Praha: Informatorium, 2007. 271 s. ISBN 80-7333-039-3.

[2] OECD. Handbook on Constructing Composite Indicators. Metodology and User Guide. Paris: Organisation for Economic Co-operation and Development, 2008. 158 s. ISBN 978-92-64-04345-9.

[3] EUROSTAT. Europe 2020 http://epp.eurostat.ec.europa.eu/portal/page/portal/europe_2020_indicators/headline_indicators

[4] FISCHER JAN, FISCHER JAKUB. EU – ostrý start statistiky nové ekonomie. *Statistika*, 2002, roč. 39, č. 11–12, s. 435–445. ISSN 0322-788X

Adresa autora:

Ing. Lenka Hudrlíková
Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra ekonomické statistiky
nám. W. Churchilla 4,
130 67 Praha 3
xhudl05@vse.cz

Příspěvek vznikl s podporou projektu IGA VŠE č.30/2010 "Odhady multifaktorové produktivity".

Využitie metódy priamej štandardizácie pri výskume plodnosti štátov sveta

The exploitation of methods of direct standardization in research fertility of the countries of the world

Andrej Chromeček

Abstract: This article presents general analysis level of fertility, which is studied by the method of direct standardization. This analyse is based on datas from Population Division of United Nations

Key words: direct method of standardisation, world population, general fertility rate, age-specific fertility rate

Kľúčové slová: priama štandardizácia, svetová populácia, všeobecná miera plodnosti, špecifická plodnosť podľa veku.

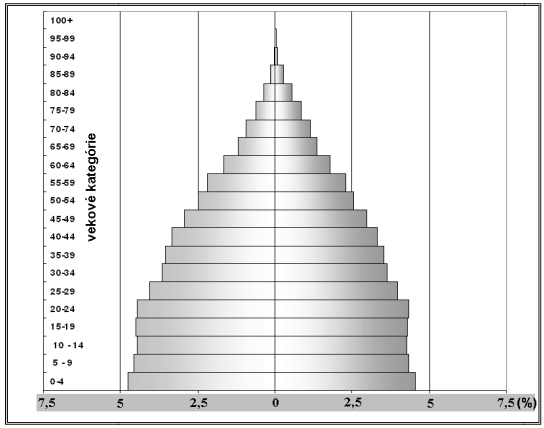
JEL clasification: Y50

1. Úvod

Intenzita výskytu väčšiny demografických udalostí je značne závislá najmä od pohlavnej a vekovej štruktúry skúmanej populácie. Pri výskume plodnosti je relatívne vhodným ukazovateľom všeobecná miera plodnosti, ktorá nám udáva koľko detí sa narodí za skúmané obdobie (najčastejšie rok) tisícke žien v reprodukčnom veku 15 až 49 rokov. Už z konštrukcie tohto ukazovateľa nám vyplýva, že je očistený od pohlavnej štruktúry, keďže berieme do úvahy len ženskú časť populácie. Čiastočne je očistená aj veková štruktúra, keďže všeobecná miera plodnosti prepočítava počet narodených detí len na ženy vo vekových skupinách, v ktorých sú schopné otehotnieť. Za reprodukčný vek žien sa bežne zaužívalo vymedzenie od pätnásteho do začiatku päťdesiateho roku života (počty pôrodov v iných vekových kategóriách nie sú štatisticky významné). V ukazovateli všeobecnej plodnosti však nie je reprodukčný vek 15 – 49 ďalej rozčlenený. Vekovo špecifická plodnosť počas takto širokého vekového intervalu pritom samozrejme nie je konštantná. Vplyv vekovej štruktúry je možné eliminovať výpočtom špecifických mier podľa veku. My sme chceli urobiť jednu komplexnú charakteristiku plodnosti štátov sveta, ktorá by bola navzájom porovnateľná, a odstraňovala by vplyv vekovej a pohlavnej štruktúry. Rozhodli sme sa preto využiť metódu priamej štandardizácie (poznámka: v nasledujúcom texte budeme často používať zjednodušený skrátený výraz štandardizácia, pričom sa však vždy bude jednať o metódu priamej štandardizácie).

2. Metodický postup

Priama štandardizácia je technika eliminujúca vplyv vekovej štruktúry na mieru určitého demografického javu v porovnávaných populáciách. Štandardom je vhodne zvolená veková štruktúra. [1] Pri porovnávaní plodnosti z viacerých dôvodov nie je najvhodnejšie používať všeobecnú mieru plodnosti. Napriek tomu, že je to presnejší ukazovateľ ako napríklad hrubá miera pôrodnosti, ktorá počty živo narodených detí vzťahuje ku strednému stavu populácie, aj všeobecná miera plodnosti má svoje nedostatky pri medzinárodnom porovnávaní. Podiel žien v reprodukčnom veku nie je úplne konštantný v každej populácii, a pokiaľ sa pozrieme na rozčlenenie tohoto samotného súboru podľa jemnejších vekových kategórií tak nájdeme ešte väčšie nuansy.



Obrázok 8: Veková štruktúra svetovej populácie, zvolená za štandard (priemer 2005 -2010)

Metóda priamej štandardizácie nám odstraňuje práve vplyv vekovej (pri nami zvolenom postupe aj pohlavnej) štruktúry, keďže skutočnú vekovú štruktúru každej populácie, zameníme za spoločný štandard. Zvyčajne sa ako štandard zvykne používať veková štruktúra hierarchicky vyššej populácie, v ktorej sú jednotlivé subpopulácie zhrnuté. Rovnaký postup sme využili aj pri tejto štandardizácii a za štandard sme zvolili vekovú štruktúru celosvetovej populácie, rozčlenenú aj podľa pohlaví. Na základe dostupnosti údajov z populačnej divízie OSN sme boli nútený pracovať s 5 ročnými vekovými kategóriami. Vstupnými dátami boli absolútne súčasné počty všetkých obyvateľov každej jednotlivej krajiny, ktoré boli vždy podľa pomerných čísel vyrátaných z celosvetovému priemeru (viď tab. 1) prerozdelené do jednotlivých vekových kategórií.

Tabuľka 1: Percentuálne podiely počtu obyvateľov podľa vekových kategórií svetovej populácie (2010)

vekové kategórie	**0-4**	**5 - 9**	**10 - 14**	**15-19**	**20-24**	**25-29**	**30-34**	**35-39**	**40-44**	**45-49**	**50-54**
muži (%)	4,889	4,715	4,686	4,695	4,462	4,039	3,782	3,603	3,288	2,874	2,494
ženy (%)	4,564	4,399	4,383	***4,426***	***4,260***	***3,895***	***3,664***	***3,509***	***3,228***	***2,849***	2,507

vekové kategórie	**55-59**	**60-64**	**65-69**	**70-74**	**75-79**	**80-84**	**85-89**	**90-94**	**95-99**	**100+**	**spolu**
muži (%)	2,056	1,556	1,209	0,920	0,615	0,338	0,138	0,042	0,009	0,001	50,412
ženy (%)	2,111	1,645	1,342	1,095	0,809	0,519	0,255	0,098	0,028	0,005	49,588

Výpočet štandardizovanej miery plodnosti si môžeme názorne ilustrovať na príklade Slovenskej republiky. Vstupným údajom do výpočtov bol priemerný počet obyvateľov za roky 2005 – 2010 z vrchnej časti tabuľky 2. s hodnotou 5 399 000. Táto hodnota bola prenásobená relatívnymi hodnotami z tabuľky č. 1 , čím sa prerozdelila do vekových kategórií podľa pohlaví rovnakým percentuálnym pomerom ako celosvetová populácia. V spodnej časti tabuľky č.2 vidíme tieto štandardizované absolútne počty obyvateľov Slovenskej republiky.

Tabulka č. 2 Veková štruktúra Slovenskej republiky

pôvodná veková štruktúra priemer za roky 2005/2010 (počet obyvateľov v tisíckach)

vek	0-4	5 - 9	10 - 14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Spolu	266	273,5	324	386,5	431,5	462	447	391,5	372,5	391	395,5
Muži	136,5	140	165,5	197,5	220	235	227,5	198	186,5	194,5	193,5
Ženy	129,5	133,5	158,5	***189***	***212***	***227***	***221***	***194***	***186***	***197***	202

vek	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100+	spolu
Spolu	347	264,5	204	169	134	89,5	36,5	10,5	2,5	0	**5399**
Muži	164,5	119	85,5	65,5	47,5	28	10,5	2,5	0,5	0	**2618**
Ženy	182	145,5	118,5	103,5	87	61,5	26	8	2	0	**2782**

Priamo štandardizovaná veková štruktúra priemer za roky 2005/2010 (počet obyvateľov v tisíckach)

vek	0-4	5 - 9	10 - 14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54
Spolu	510,3	492,0	489,6	492,4	470,8	428,3	401,9	383,9	351,8	308,9	270,0
Muži	263,9	254,5	253,0	253,5	240,9	218,1	204,1	194,5	177,5	155,1	134,7
Ženy	246,4	237,5	236,6	***238,9***	***230,0***	***210,3***	***197,8***	***189,4***	***174,2***	***153,8***	135,4

vek	55-59	60-64	65-69	70-74	75-79	80-84	85-89	90-94	95-99	100+	spolu
Spolu	225,0	172,8	137,7	108,8	76,9	46,3	21,2	7,5	2,0	0,3	**5399**
Muži	111,0	84,0	65,3	49,7	33,2	18,2	7,4	2,3	0,5	0,1	**2721**
Ženy	113,9	88,8	72,4	59,1	43,7	28,0	13,7	5,3	1,5	0,3	**2677**

Na prvý pohľad je vidno, že štandardizovaná pohlavná štruktúra Slovenskej Republiky by bola mierne modifikovaná od pôvodnej. Zatiaľ čo u nás podobne ako vo všetkých rozvinutých krajinách máme pozitívnu femininitu - čiže prevahu žien, v svetovej populácii, ktorá je tvorená väčšinou rozvojovými krajinami je prevaha mužov, čo vidno na štandardizovanej štruktúre v spodnej časti tabuľky č. 2. Ešte väčšie zmeny ako pohlavná štruktúra by zaznamenala veková štruktúra. Pri výpočte špecifickej plodnosti sú pre nás dôležité len počty žien vo vekových kategóriách 15 – 49 rokov. Pri porovnaní je vidieť, že by došlo k nárastu počtu žien v najnižších vekových kategóriách fertilného veku 15 – 25 rokov, na druhej strane vo vyšších vekových kategóriách 25 -49 by počty žien poklesli. Štandardizované počty žien vo veku 15 -49 rokov sme následne prenásobili originálnymi hodnotami vekovo špecifických plodností (viď. tabuľka č.3), čím sme dostali štandardizované počty živonarodených detí. Po ich zosumovaní sa ukázalo, že ročne by sa na Slovensku pri zachovaní vekovo špecifických plodností ale pri zmene vekovo pohlavnej štruktúry narodilo o cca 900 detí menej oproti nezmenenému stavu.

Tabuľka č. 3. Plodnosť v Slovenskej republike

vek	15-19	20-24	25-29	30-34	35-39	40-44	45-49	spolu
originálna špecifická plodnosť podľa veku (promile)	20,7	58,02	86,78	63,67	22,54	3,68	0,2	/
originálny počet živonarodených (tisícky)	3,9	12,3	19,7	14,0	4,4	0,7	0,0	**55,0**
štandardizovaný počet živonarodených (tisícky)	4,9	13,3	18,2	12,6	4,3	0,6	0,0	**54,1**

Štandardizovanú všeobecnú mieru plodnosti sme následne jednoducho dorátali ako podiel štandardizovaného počtu všetkých detí pripadajúcich na 1000 žien štandardizovanej populácie. Rovnakým postupom, boli vyrátané hodnoty priamo štandardizovanej plodnosti pre súbor 196 krajín a závislých území s počtom obyvateľov vyšším ako 100 000, za ktoré boli dostupné dáta z populačnej divízie organizácie spojených národov.

3. Výsledky

Po tom ako sme vyrátali priamo štandardizované hodnoty za všetkých 196 populácií, pristúpili sme k ich porovnaniu s reálnymi súčasnými hodnotami. Pri 78 populáciách by zmena vekovo pohlavnej štruktúry mala za následok zvýšenie všeobecnej miery plodnosti. Naopak pri 118 populáciách vyšli hodnoty štandardizovanej všeobecnej miery plodnosti nižšie oproti skutočným hodnotám (viď. obrázok č.2.). Najvyššiu kladnú odchýlku, zaznamenalo závislé územie v karibiku: Holandské Antily, kde priamo štandardizovaná miera všeobecnej plodnosti bola až o 17,4‰ vyššia ako reálne hodnoty. Naopak najväčší prepad tohto ukazovateľa o -26,7‰ by zaznamenal africký štát Niger. Vo všeobecnosti sa dá tvrdiť, že ekonomicky a demograficky rozvinuté krajiny by zmenou pohlavno-vekovej štruktúry zvýšili svoju mieru všeobecnej plodnosti, zatiaľ čo u najmenej rozvinutých krajín sveta by došlo k poklesu tohto ukazovateľa. Avšak nie vždy sa toto pravidlo potvrdzuje. Aj u niektorých z demograficky vyspelých krajín by došlo k poklesu miery plodnosti, pokiaľ by sa veková a pohlavná štruktúra ich populácie prerozdelila do jednotlivých kategórií podľa celosvetového priemeru. Príklady takýchto krajín sú Španielsko, Írsko, Poľsko Česká Republika, alebo Japonsko. V hospodársky a demograficky rozvinutých krajinách dochádza v kontexte zmien súvisiacich s druhým demografickým prechodom k odkladaniu pôrodov do vyšších vekových kategórií, takže ženy dosahujú maximálnu špecifickú plodnosť vo vekových kategóriách 25 – 35 rokov. Keďže vo vekovej štruktúre sveta ktorú sme zvolili za štandard tvoria ženy vo vekových kategóriách 25 - 30, a 30 - 35 omnoho menší podiel ako je tomu dnes u mnohých rozvinutých krajín, ktoré majú maximá plodnosti práve v týchto vekových kategóriách, dochádza napriek relatívnemu „omladeniu“ iba k miernemu nárastu, či u niektorých krajín dokonca k poklesu všeobecnej miery plodnosti prepočítanej na všetky ženy v reprodukčnom veku. Napríklad pokiaľ by Slovenská republika mala vekovú štruktúru zhodnú so svetovou, pri zachovaní reálnych súčasných špecifických plodností podľa veku, tak miera všeobecnej plodnosti by stúpla iba o 0,2‰. Štáty v ktorých je veľká časť reprodukčného správania realizovaná vo vyšších vekových kategóriách, by iba málo profitovali so zvýšeného počtu žien v nižších vekových kategóriách (do 25rokov), ale naopak by strácali zúžením vekovej pyramídy vo vrchnej časti reprodukčného spektra (25 až 49rokov).

4. Záver

Využitie metódy priamej štandardizácie pri porovnávaní všeobecnej miery plodnosti sa ukázalo ako vhodný nástroj pre komparáciu na svetovej úrovni. Interpretácia získaných výsledkov je však často zložitá, pretože najvyššie zistené odchýlky priamo štandardizovaných hodnôt od pôvodných sa často vyskytujú v populáciách s atypickou pohlavnou, či vekovou štruktúrou, ovplyvnenou buďto mechanickým pohybom, alebo priamymi stratami spôsobenými vnútornými či medzinárodnými vojnovými konfliktmi. Bez hlbšej analýzy jednotlivých populácií je možné na tieto najextrémnejšie odchýlky poukázať, ale nie vždy ich dokážeme okamžite uspokojivo vysvetliť. Metóda priamej štandardizácie nám eliminuje vplyv vekovej štruktúry v jednotlivých populáciách, na druhej strane nám pomáha odhaliť neskreslenú úroveň všeobecnej plodnosti.

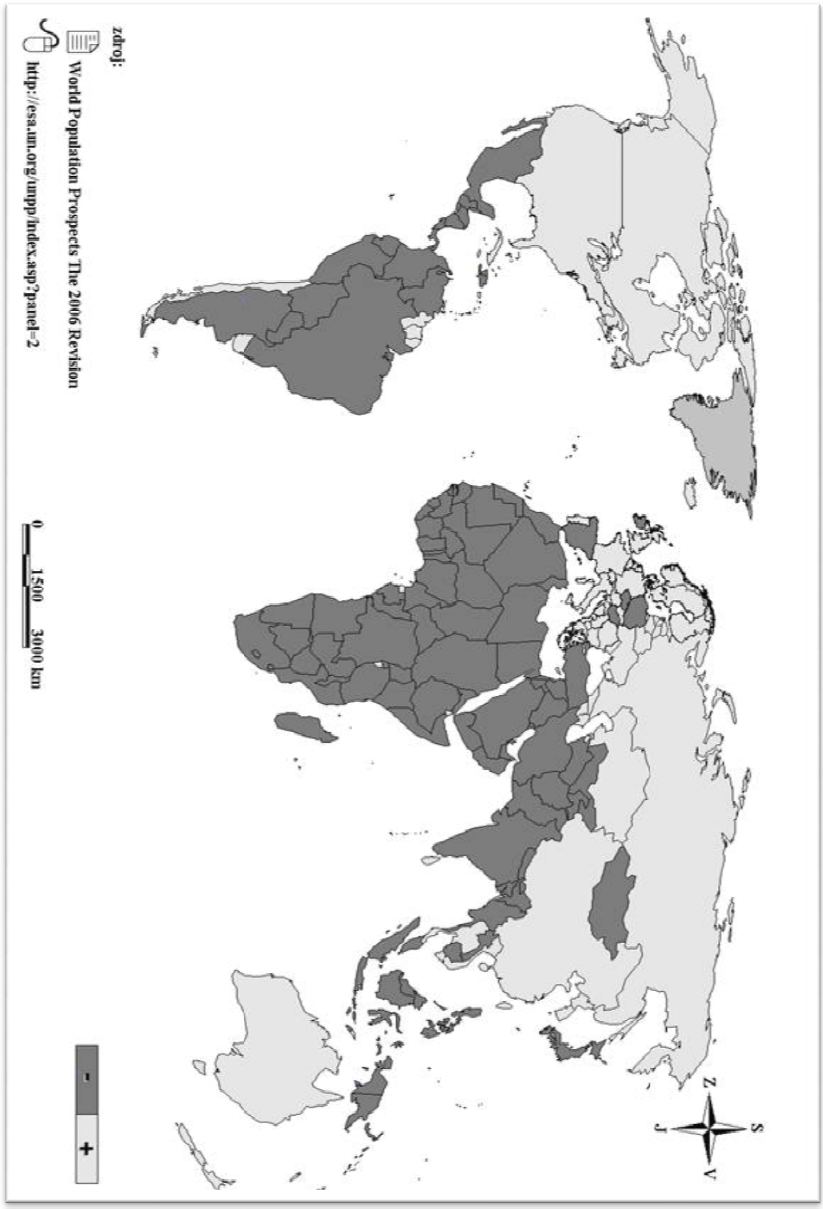
5.Literatúra

[1] JURČOVÁ, D. 2005, Slovník Demografických pojmov, Infostat. ISBN 50-85659-40-9
[2] World Population Prospects The 2006 Revision, http://esa.un.org/unpp/index.asp?panel=2

Adresa autora:

Andrej Chromecek, Mgr., Prírodovedecká fakulta UK, Mlynská dolina, Bratislava, chromecek@fns.uniba.sk

Poďakovanie: *Článok vznikol s prispením grantu UK142/2010*



Obrázok 2: Rozdiel hodnôt všeobecnej plodnosti a priamo štandardizovanej všeobecnej plodnosti

Determination of two-sided tolerance interval in a linear regression model
Výpočet obojstranných tolerančných intervalov v lineárnom regresnom modeli

Martina Chvosteková

Abstract: This article deals with methods for computing two-sided tolerance intervals in a linear regression. For non-simultaneous two-sided tolerance intervals we present also a formula for computing equal-tailed intervals. A method for computing simultaneous tolerance intervals and simultaneous equal-tailed intervals based on the confidence-set approach, where we consider a confidence-set constructed using the likelihood ratio test for testing null hypothesis about all parameters of the regression model, is introduced.

Abstrakt: Príspevok je zameraný na výpočet obojstranných tolerančných intervalov pre lineárny regresný model s normálne rozdelenými nezávislými chybami. Uvažujeme dva typy obojstranného tolerančného intervalu, jeden v literatúre prednostne nazývaný obojstranný a druhý, tzv. centrálny ("equal-tailed"). Uvádzame postupy na výpočet oboch typov pre prípad nesimultánnych intervalov tj. pre pevnú hodnotu vysvetľujúcej premennej a pre prípad simultánnych intervalov, tj. ľubovoľná hodnota vysvetľujúcej premennej. Na stanovenie simultánnych tolerančných intervalov uvádzame metódy založené na tzv. confidence-set prístupe, kde použitá oblasť spoľahlivosti pre parametre modelu bola skonštruovaná pomocou testu pomerom vierohodnosti pre testovanie nulovej hypotézy o všetkých parametroch regresného modelu zároveň.

Key words: linear regression model, tolerance interval, tolerance factor, pivot.

Kľúčové slová: lineárny regresný model, tolerančný interval, tolerančný faktor, pivot.

JEL classification: C10, C30

1. Introduction

Linear regression model is used to express statistical dependence of a response variable and a predictor in various disciplines of science. In particular, we suppose that a relationship between the response variable and the predictor is described by a multiple linear regression model, where we also assume that the response variables are normally distributed with a common variance σ^2. The linear regression model is represented in the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{Z}, \tag{1}$$

where $\mathbf{Y} = (Y_1,\ldots,Y_n)^T$ represents an $n\times 1$ random vector of independent observations, $\mathbf{X}$ is an $n\times q$ $(n > q)$ matrix of rank q, with constant elements. Vector $\boldsymbol{\beta} = (\beta_0,\beta_1,\ldots,\beta_{q-1})^T$ and standard deviation $\sigma > 0$ represent unknown parameters of the regression model and $\mathbf{Z}$ is an $n\times 1$ vector of standard normal errors, i.e. $\mathbf{Z} \sim N(\mathbf{0},\mathbf{I}_n)$. Under the assumptions, the least squares estimators of $\boldsymbol{\beta},\sigma$ are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad \text{and} \qquad S^2 = \frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{n-q}. \tag{2}$$

Note that $\hat{\boldsymbol{\beta}} \sim N_q(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ and $(n-q)S^2/\sigma^2 \sim \chi^2_{n-q}$, where χ^2_{n-q} denotes a central chi-square random variable with $n-q$ degrees of freedom. Random variables $\hat{\boldsymbol{\beta}}$ and S^2 are independent.

A tolerance interval is specified by its *content* and *confidence level,* denoted $0<\gamma<1$ and $0<1-\alpha<1$, respectively. Throughout this paper we will refer simply to $(\gamma,1-\alpha)$ tolerance interval. In practical applications the values are close to one. A future observation of a response at the predictor $\mathbf{x}^T=(1,x_1,\ldots,x_{q-1})^T$ is written as $Y(\mathbf{x})=\mathbf{x}^T\boldsymbol{\beta}+\sigma Z$, where $Z \sim N(0,1)$ and $Y(\mathbf{x})$ is assumed to be independent of $\mathbf{Y}$.

For a fixed predictor $\mathbf{x}$, a $(\gamma,1-\alpha)$ two-sided tolerance interval for a future observation $Y(\mathbf{x})$ is considered in the following (general) form

$$\left\langle \mathbf{x}^T\hat{\boldsymbol{\beta}}-\lambda(\mathbf{x}\,|\,\gamma,1-\alpha,\mathbf{X})S, \mathbf{x}^T\hat{\boldsymbol{\beta}}+\lambda(\mathbf{x}\,|\,\gamma,1-\alpha,\mathbf{X})S \right\rangle, \tag{3}$$

where $\lambda(\mathbf{x}\,|\,\gamma,1-\alpha,\mathbf{X})$ is a *tolerance factor* for the given content γ, confidence level $1-\alpha$ and $\mathbf{X}$, we will use for simplification a shorter notation $\lambda(\mathbf{x})$.

Non-simultaneous (with a fixed $\mathbf{x}$) $(\gamma,1-\alpha)$ two-sided tolerance interval (NTTI) for $Y(\mathbf{x})$ is an interval of the form (3) that contains at least a γ proportion of the distribution random variable $Y(\mathbf{x})$ with confidence level $1-\alpha$ constructed using the vector of observations $\mathbf{Y}$. The tolerance factor is to be determined subject to the content and confidence level requirements. There is a second type of non-simultaneous two-sided tolerance interval, which is constructed to contain at least a γ proportion of the center of the $Y(\mathbf{x})$-distribution with confidence level $1-\alpha$ (CNTTI). The exact values of tolerance factors for NTTI and CNTTI are obtained by numerically solving the integral equations.

The simultaneous $(\gamma,1-\alpha)$ two-sided tolerance intervals (STTI) are constructed using the vector of observations $\mathbf{Y}$, so that with the confidence level $1-\alpha$, at least a γ proportion of the $Y(\mathbf{x})$-distribution is to be contained in the corresponding tolerance interval, simultaneously for all possible values of predictors $\mathbf{x}$. There is no procedure to compute the tolerance factor $\lambda(\mathbf{x})$ for STTI satisfying the definition exactly. Lieberman and Miller (SW) in [5] were the first who considered the problem of finding a simultaneous tolerance intervals in a linear regression model. They presented an approximation for the case of a simple linear regres-sion. Further methods for computing STTI in a linear regression, Wilson (W) [10], Limam-Thomas (PS) [6] and modified Wilson (MW) [6] and LRT method (LRT) [3], are based on confidence-set approach. Mee-Eberhardt-Reeve [7] obtained the narrowest tolerance intervals, but they considered the STTI for limited range of possible values of preditors. The simultaneous equal-tailed tolerance intervals (CSTTI) are constructed to covering at least a γ proportion of the center of the $Y(\mathbf{x})$-distribution simultaneously for any predictor $\mathbf{x}$, with confidence level $1-\alpha$. Scheffé obtained narrower intervals than those of Lieberman-Miller (CSx) presented in [5], but he restricted the range of possible values of $\mathbf{x}$. In [3] there is suggested a procedure (CLRT) for computing CSTTI based on confidence-set approach.

In section 2 we describe classical approach to determination of tolerance intervals in a linear regression model. The result for computing NTTI is given and we present the result for computing CNTTI derived by applying the procedure for computing an interval for univariate normal distribution. In section 3 we describe the confidence set approach for computing STTI, also CSTTI, and introduce the methods based on the exact likelihood ratio test for testing null hypothesis about all parameters of the model [1]. In section 4 the tolerance factors for NTTI, CNTTI , for STTI by W, LT, MW, LRT, for CSTTI by CLRT, are computing in an

example. The results are compared for a given $\gamma = 0.95$, $\alpha = \{0.05, 0.01\}$ level at several specified predictors.

2. Non-simultaneous two-sided tolerance intervals

Here we present classical approach to determination of a tolerance interval in a linear regression model, where pivotal quantities are used. The results for computing NTTI and CNTTI are given.

Let $C(\mathbf{x};\hat{\boldsymbol{\beta}},S) = P_{Y(\mathbf{x})}(\mathbf{x}^T\hat{\boldsymbol{\beta}} - \lambda(\mathbf{x})S \le Y(\mathbf{x}) \le \mathbf{x}^T\hat{\boldsymbol{\beta}} + \lambda(\mathbf{x})S \mid \hat{\boldsymbol{\beta}}, S)$ denote the content for the tolerance interval (3), given $\hat{\boldsymbol{\beta}}$ and S. Denote

$$\mathbf{b} = \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma} \sim N(0, (\mathbf{X}^T\mathbf{X})^{-1}) \quad \text{and} \quad u = \frac{S}{\sigma},\ (n-q)u^2 \sim \chi^2_{n-q}. \tag{4}$$

Random variables $\mathbf{b}, u$ are independently distributed and their distributions are free of unknown parameters $\boldsymbol{\beta}, \sigma$, dependent only on the matrix $\mathbf{X}$. Now, we can rewrite the content,

$$C(\mathbf{x};\hat{\boldsymbol{\beta}},S) = \Phi(\mathbf{x}^T\mathbf{b} + \lambda(\mathbf{x})u) - \Phi(\mathbf{x}^T\mathbf{b} - \lambda(\mathbf{x})u) = C(\mathbf{x};\mathbf{b},u), \tag{5}$$

where Φ denotes cumulative distribution function of the standard normal distribution.

NTTI are constructed using the vector of observations $\mathbf{Y}$ for the given matrix $\mathbf{X}$ so that at least a γ proportion of distribution of $Y(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta} + \sigma Z \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2)$ is contained with confidence $1-\alpha$, hence

$$P_{\mathbf{b},u}(C(\mathbf{x};\mathbf{b},u) \ge \gamma) = 1 - \alpha. \tag{6}$$

For a fixed predictor $\mathbf{x}$, NTTI are actually similar to a two-sided tolerance interval for a univariate normal distribution with unknown mean and variance. Hence tolerance factor satisfying the condition (6) is obtained by applying the result for univariate normal distribution presented in [4]. Tolerance factor $\lambda = \lambda(\mathbf{x})$ satisfying (6) is solution of the integral equation

$$\sqrt{\frac{2}{\pi d^2}} \int_0^\infty P\left(\chi^2_{n-q} > \frac{(n-q)\chi^2_{1,\gamma}(x^2)}{\lambda^2}\right) e^{-\frac{x^2}{2d^2}} dx = 1 - \alpha, \tag{7}$$

where $d^2 = \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}$, $\chi^2_{1,\gamma}(x^2)$ denotes the γ–quantile of the non-central chi-square distribution with one degree of freedom and non-centrality parameter x^2. Wallis [9] proposed an analytic approximation $\lambda_W = \sqrt{(n-q)\chi^2_{1,\gamma}(d^2)/\chi^2_{n-q}(\alpha)}$, where $\chi^2_{n-q}(\alpha)$ denotes the α–quantile of chi-square distribution with $n-q$ degrees of freedom.

There is another type of NTTI, which is usually referred to as an equal-tailed tolerance interval. A $(\gamma, 1-\alpha)$ equal-tailed tolerance interval is constructed so that, with $1-\alpha$ confidence no more than a proportion $(1-\gamma)/2$ of the $Y(\mathbf{x})$-distribution is greater than $\mathbf{x}^T\hat{\boldsymbol{\beta}} + \lambda_C S$ and no more than a proportion $(1-\gamma)/2$ of the population is less than $\mathbf{x}^T\hat{\boldsymbol{\beta}} - \lambda_C S$. In terms of pivotal quantities $\mathbf{b}$, u, the condition should be satisfied is

$$P_{\mathbf{b},u}\left(\mathbf{x}^T\mathbf{b} - \lambda_C u \le -u_{(1+\gamma)/2} \quad \wedge \quad \mathbf{x}^T\mathbf{b} + \lambda_C u \ge u_{(1+\gamma)/2}\right) = 1 - \alpha. \tag{8}$$

Using analogical procedure as was used for deriving the equal-tailed interval for univariate normal distribution, we obtained that tolerance factor λ_C is the solution of the integral equation

$$\left[2^{\frac{n-q}{2}}\Gamma\left(\frac{n-q}{2}\right)\right]^{-1}\int_{\left[u_{(1+\gamma)/2}/\lambda_C\right]^2(n-q)}^{\infty}\left(2\Phi\left(\frac{-u_{(1+\gamma)/2}+\lambda_C\sqrt{x/(n-q)}}{d}\right)-1\right)e^{-x/2}x^{\frac{n-q}{2}-1}dx=1-\alpha, \quad (9)$$

where Γ denotes the gamma function. To compute λ_C, an initial value for a numerical integration procedure is set as $d\times t_{n-q,1-\alpha}\left(u_{(1+\gamma)/2}/d\right)$, where $t_{n-q,1-\alpha}\left(u_{(1+\gamma)/2}/d\right)$ denote the $(1-\alpha)$ – quantile of the non-central t distribution with $n-q$ degrees of freedom with parameter of non-centrality $u_{(1+\gamma)/2}/d$.

3. The simultaneous tolerance interval

In this section, the confidence-set approach and the likelihood ratio test for testing null hypothesis about all parameters of the regression model with normally distributed errors are described. We introduce the LRT method suggested in [3] for computing STTI and the method for computing CSTTI suggested in [1], which are based on the confidence-set approach, where the confidence-set is constructed using the likelihood ratio test.

The simultaneous $(\gamma,1-\alpha)$ two-sided tolerance intervals of the form (3) in a linear regression model with normally distributed errors are constructed using vectors of observations $\mathbf{Y}$ such that, with the confidence level $1-\alpha$, at least the specified proportion γ of the $Y(\mathbf{x})=\mathbf{x}^T\boldsymbol{\beta}+\sigma Z$ distribution is to be contained in the corresponding interval, simultaneously for all $\mathbf{x}\in\mathsf{R}^q$, where $\boldsymbol{\beta},\sigma$ are unknown parameters of the model, $Z\sim N(0,1)$.
The condition to be satisfied is

$$P_{\mathbf{b},u}(C(\mathbf{x};\mathbf{b},u)\geq\gamma \quad \forall\mathbf{x}\in\mathsf{R}^{q\times 1})=1-\alpha. \quad (10)$$

The function $\lambda(\mathbf{x})$ that satisfies (10) is called a simultaneous tolerance factor. Denote the set $G(\mathbf{X})$ of $\mathbf{b},u$ satisfying the equation (9) for a given matrix $\mathbf{X}$; it is called a $(1-\alpha)$-pivotal set. A confidence set for both true $\boldsymbol{\beta}$, σ can be described in terms of the $(1-\alpha)$-pivotal set $G(\mathbf{X})$, $C_{1-\alpha}(\mathbf{Y}\mid\mathbf{X})=\{(\boldsymbol{\beta},\sigma)=(\hat{\boldsymbol{\beta}}-\mathbf{b}S/u,S/u):(\mathbf{b},u)\in G(\mathbf{X})\}$. The confidence-set approach for constructing STTI consists in defining the set $G(\mathbf{X})$, and then the simultaneous tolerance factor based on it satisfies

$$\lambda(\mathbf{x})=\min\{\lambda: C(\mathbf{x};\mathbf{b},u)\geq\gamma \quad (\mathbf{b},u)\in G(\mathbf{X})\}. \quad (11)$$

In [2] there is presented a brief overview of methods for computing STTI based on the confidence-set approach, Wilson [10], Limam-Thomas [6], modified Wilson [6]. Pivotal sets proposed in these method are approximately $(1-\alpha)$ and all the STTI are conservative.

The test statistic for testing null hypothesis about all parameters of the regression model (1), i.e. $H_0:(\boldsymbol{\beta},\sigma)=(\boldsymbol{\beta}_0,\sigma_0)$ against $H_1:(\boldsymbol{\beta},\sigma)\neq(\boldsymbol{\beta}_0,\sigma_0)$, is given by

$$\mathsf{D}(\mathbf{Y}|\mathbf{X})=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0)/\sigma_0^2-n\ln\left(^2(n-q)S^2/n\sigma_0^2\right)-n. \quad (12)$$

The distribution $\mathsf{D}(\mathbf{Y}|\mathbf{X})$ depends only on the number of observations n and on number of predictors covariates q. The critical values denoted $\mathsf{D}_{1-\alpha}(n,q)$ for testing null hypothesis $H_0:(\boldsymbol{\beta},\sigma)=(\boldsymbol{\beta}_0,\sigma_0)$ are presented in [1] for different values of $q=1,\ldots,10$ and selected sample sizes, $n=q+1:(1):40$, $n=45:(5):100$ and ∞, and for usual significance levels $\alpha=\{0.1,0.05,0.01\}$. The exact $(1-\alpha)$ confidence-set for $\boldsymbol{\beta}$, σ derived in [1] by using the test is given by

$$C_{1-\alpha}(\mathbf{Y}|\mathbf{X})=\left\{(\boldsymbol{\beta},\sigma):\mathsf{D}(\mathbf{Y}|\mathbf{X})\leq\mathsf{D}_{1-\alpha}(n,q)\right\}. \quad (13)$$

The LRT method suggested in [3] is based on the confidence-set approach combination with a part of Wilson procedure, the pivotal set is constructed using the exact likelihood ratio test for testing null hypothesis about all parameters of the regression model. The exact $(1-\alpha)$ pivotal set is given as

$$G_{LRT}(\mathbf{X}) = \left\{(\mathbf{b},u) : u^2(n-q) + \mathbf{b}^T(\mathbf{X}^T\mathbf{X})\mathbf{b} - n\ln(u^2(n-q)) + n(\ln(n)-1) \le \mathsf{D}_{1-\alpha}(n,q)\right\}$$

(14) The simultaneous tolerance factor for predictor $\mathbf{x}$ is the greater solution of the non-linear equation

$$(r-r_0)^2 - h^2 - n\delta^2(\mathbf{x}) - n\delta^2(\mathbf{x})\ln\left(r^2(n-q)/n\lambda^2\right) + \delta^2(\mathbf{x})r^2(n-q)/\lambda^2 - \mathsf{D}_{1-\alpha}\delta^2(\mathbf{x}) = 0, \quad (15)$$

where $r_0 = \Phi^{-1}(\gamma)$. For certain selected γ, Wilson tabulated values of h^2, for example $h^2 = 0.25$ and $h^2 = 0.0438$ for $\gamma = 0.75$ and $\gamma = 0.95$, respectively. We consider the equation for (15) $n, \lambda^2, r^2, \delta^2(\mathbf{x}) \neq 0, n > q$, and with restriction $r, \lambda \in (0,\infty)$.

The simultaneous equal-tailed tolerance intervals (CSTTI) are constructed so that, with confidence level $1-\alpha$ no more than $(1-\gamma)/2$ of the distribution of the $Y(\mathbf{x})$-distribution is greater than $\mathbf{x}^T\boldsymbol{\beta} + \lambda_{CS}(\mathbf{x})S$ and no more than a proportion $(1-\gamma)/2$ of the population is less than $\mathbf{x}^T\boldsymbol{\beta} - \lambda_{CS}(\mathbf{x})S$, for any predictor $\mathbf{x}$. The simultaneous central tolerance factor $\lambda_{CS}(\mathbf{x})$ based on $C_{1-\alpha}^{LRT}(\mathbf{Y}|\mathbf{X})$ is computed in the procedure CLRT suggested in [1] as

$$\lambda_{CS}(\mathbf{x}) = 1/S\left(\mathbf{x}^T\boldsymbol{\beta} - \inf_{(\boldsymbol{\beta},\sigma)\in C_{1-\alpha}(\mathbf{Y}|\mathbf{X})}\left\{\mathbf{x}^T\boldsymbol{\beta} + u_{(1-\gamma)/2}\sigma\right\}\right) = 1/S\left(\sup_{(\boldsymbol{\beta},\sigma)\in C_{1-\alpha}(\mathbf{Y}|\mathbf{X})}\left\{\mathbf{x}^T\boldsymbol{\beta} + u_{(1+\gamma)/2}\sigma\right\} - \mathbf{x}^T\boldsymbol{\beta}\right). \quad (16)$$

4. Discussion

The tolerance factors for non-simultaneous (NTTI, CNTTI) and the simultaneous tolerance intervals (W, PS, MW, LRT, CLRT) are computed in an example. The results are compared for a given $\gamma = 0.95$, $\alpha = \{0.05, 0.01\}$ level at several specified predictors.

A simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \sigma Z_i$, with independent errors $Z_i \sim N(0,1)$, $i = 1,\ldots,15$ (i.e. n = 15 and q = 2) was considered for the analysis of the orifice-speed measurements. For details, refer to the paper [5]. Table 1 includes values of the tolerance factors computed for two combinations of $(\gamma, 1-\alpha)$, and, for each combination, the same four values of predictors were used $x^* = |x - \bar{x}|/S = \{0,1,2,3\}$. The first row states values are obtained for non-simultaneous, the second for central non-simultaneous. The simultaneous tolerance factors are computed by the all methods based on confidence-set approach. The tolerance factors for CSTTI are computed by the CLRT.

Table 6: Tolerance factors for TTI in the linear regression model for $\gamma = 0.95$ at chosen predictors and confidence level $\alpha = \{0.05, 0.01\}$.

	α = 0.01					***α = 0.05***			
x^*	***0***	***1***	***2***	***3***		***0***	***1***	***2***	***3***
NS	***1.37***	***1.39***	***1.42***	***1.45***		***1.53***	***1.56***	***1.60***	***1.64***
CNS	***1.44***	***1.47***	***1.51***	***1.54***		***1.62***	***1.65***	***1.71***	***1.77***
W	***5.81***	***5.94***	***6.60***	***7.55***		***4.30***	***4.43***	***4.95***	***5.68***
MW	***5.01***	***5.15***	***5.74***	***6.58***		***3.85***	***3.97***	***4.43***	***5.07***
PS	***4.24***	***4.66***	***5.55***	***6.55***		***3.52***	***3.84***	***4.52***	***5.31***

LRT	***4.04***	***4.34***	***5.10***	***6.04***		***3.38***	***3.60***	***4.18***	***4.91***
CLRT	***4.54***	***4.87***	***5.62***	***6.55***		***3.80***	***4.05***	***4.63***	***5.35***

As we expected, the equal-tailed tolerance intervals are wider than their non equal-tailed counterparts, both in case of simultaneous and non-simultaneous intervals. The simultaneous intervals are wider than the non-simultaneous in general. Based on the results presented in Table 1, we recommend LRT method for computing STTI.

5. Acknowledgement

The paper was supported by the Scientific Grant Agency of the Slovak Republic (VEGA), grant 1/0077/09 and 2/0019/10.

6. References

[1]CHVOSTEKOVÁ, M., WITKOVSKÝ, V. 2009. Exact Likelihood Ratio Test for the Parameters of the Linear Regression Model with Normal Errors. In: Measurement Science Review, č. 1, 2009, s. 1 - 8.

[2]CHVOSTEKOVÁ, M. 2010: Simultánne obojstranné tolerančné intervaly v lineárnom regresnom modeli. INFORMAČNÍ BULLETIN České statistické společnosti, č. 3, 2010, s. 57 - 64.

[3]CHVOSTEKOVÁ, M. 2010: Stanovenie simultánnych obojstranných tolerančných inter--valov v lineárnom modeli založené na oblasti spoľahlivosti parametrov modelu. In: FORUM STATISTICUM SLOVACUM, č. 4, 2010, s. 73 - 78.

[4]KRISHNAMOORTHY, K. - MATHEW, T. 2009. Statistical Tolerance Regions: Theory, Applications, and Computation, Wiley, 2009. 484 s. ISBN 978-0-470-38026-0

[5]LIEBERMAN, G. J. - MILLER, R. G., Jr. 1963. Simultaneous Tolerance Intervals in Regression. In: Biometrika, č. 1/2, 1963, s. 155 - 168.

[6]LIMAM, M. M. T. - THOMAS, R. 1988. Simultaneous Tolerance Intervals for the Linear Regression Model. In: Journal of the American Statistical Association, č. 403, 1988, s. 801 - 804.

[7]MEE, R. W. - EBERHARDT, K. R. - REEVE, C. P. 1991 Calibration and Simultaneous Tolerance Intervals for Regression. In: Technometrics, č.2, 1991, s. 211-219.

[8]SCHEFFÉ, H. 1973 A Statistical Theory of Calibration. In: The Annals of Statistics, Vol. 1, č. 1, s. 1 - 37.

[9]WALLIS, W. A. 1951. Tolerance Intervals for Linear Regression, In: SECOND BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTOCS AND PROBABILITY. 1951, University of California, s. 43 - 51.

[10]WILSON, A. L. 1967. An Approach to Simultaneous Tolerance Intervals in Regression. In: The Annals of Mathematical Statistics, č. 38, 1967, s. 1536 - 1540.

Adresa autora:

Martina Chvosteková, Mgr.
Ústav merania SAV
Dúbravská cesta 9
841 04 Bratislava
chvosta@gmail.com

Základné inovačné trendy v rozvoji manažmentu a diferencie medzi druhmi/generáciami manažmentu II.

Basic Innovative Trends in Development of Management and Differences between Types/Generations of Management II.

Ľubomír Jemala

Abstract: Managerial innovation creates a starting point throughout each innovation process. In particular, Strategic management is awaiting a radical change within the next decade. This change should/must respect a trajectory of development in the subsequent line: past - present - future. This contribution attempts to generally identify selected attributes and trends of management development for about hundred years ago and by the support of the selected experts to outline the possible way forward in the context of world projections for the nearest decade.

Key words: Industrial management, Knowledge management, Management of first-generation [M 1.0], Management of second-generation [M 2.0].

Kľúčové slová: industriálny manažment, poznatkový manažment, manažment prvej generácie [M 1.0], manažment druhej generácie [M 2.0].

JEL classification: M10, M31, O30, O31.

1. Namiesto úvodu

V tomto príspevku nadväzujeme druhou časťou na prvú časť príspevku, ktorý bol publikovaný vo FSS č.4/2010. Sústreďujeme sa tu na ďalšie trendy a možné postupy v rozvoji a inováciách manažmentu, ako aj na konfrontáciu generácií manažmentu smerom k tvorbe jeho druhej generácie (M 2.0), ako ju avizuje najmä G. Hamel [1].

2. Ďalšie vývojové trendy v rozvoji manažmentu v kontexte predpovedí expertov v oblasti súradníc vývoja sveta do roku 2025

Zdroje poznatkov, použité pramene, pomocou ktorých sme spracovali širšie poňatý náčrt dlhodobých predpovedí možného vývoja našej planéty a teda nepriamo aj navigáciu trendov rozvoja strategického manažmentu, uvádzame na konci tohto príspevku v zozname literatúry. Na tomto mieste možno spomenúť len stručný výber týchto zaujímavých predpovedí.

Medzinárodné globálne/multipolárne rozloženie síl. Do r. 2025 sa má takto zmeniť zo súčasného stabilnejšieho bipolárneho rozloženia. Súčasne sa má zvýšiť vplyv neštátnych subjektov (nadnárodné obchodné spoločnosti, náboženské organizácie či aj zločinecké siete!). USA údajne majú zostať nezávislou, najsilnejšou veľmocou, ale už s menej dominantným postavením.

Vzniknú nové výrazné hrozby/riziká, tlaky. Zosilnieť má napr. boj o trhy, investície, technologické inovácie, zisky atď.

Rozvoj multilaterálnej (mnohostrannej) kooperácie. Politikov i lídrov k tomuto má donútiť predovšetkým proces transformácie medzinárodného rozloženia síl.

Ekonomický rast posilní rolu nových hráčov. Pôjde okrem krajín BRIC (Brazília, Rusko, India a Čína) aj o Indonéziu, Irán, Turecko atď. Iné krajiny budú aj naďalej ekonomicky zaostávať. Subsaharská Afrika má zostať najzraniteľnejším regiónom.

Očakáva sa nárast objemu, rýchlosti a smeru presunu globálneho bohatstva a ekonomickej sily, čo asi nemá obdobu v modernej histórii... Spôsobili to najmä dva faktory. *Rast cien ropy a komodít*, ktorý priniesol neočakávané finančné príjmy štátom okolo Perzského zálivu či Rusku. *Spracovateľský priemysel sa presunul do ázijských krajín* aj vďaka nižším nákladom a ústretovej zahraničnej politike.

Pravdepodobný je rozvoj celkom odlišného modelu riadenia „štátneho kapitalizmu". Štáty BRIC pri svojom rozvoji teda namajú/nebudú uplatňovať liberálny model západných štátov. Ide o **systém riadenia ekonomiky, v ktorom hlavnú úlohu hrá štát.** Používajú ho napr. iné rastúce ekonomiky ako Južná Kórea, Taiwan či Singapur.

V politickom živote sa očakáva nárast demokratizácie – aj v Rusku, Číne a inde. Očakávajú sa štyri strategické scenáre (podrobnejšie pozri uvedený zdroj poznania).

O celkový rast obyvateľstva sa v priebehu 20 rokov majú najviac pričiniť Ázia, Afrika a Latinská Amerika. Len 3% z celkového prirodzeného prírastku obyvateľstva planéty majú pripadnúť na Západ.

Bohatstvo na hlavu obyvateľa má byť v Európe a Japonsku stále vyššie ako v novo vznikajúcich veľmociach akými sú Čína a India ?!

Počet ekonomicky aktívneho obyvateľstva sa má v Európe aj v Japonsku znižovať. Nevyhnutné bude vyvinúť enormné úsilie na udržanie rastu!

Strategické zdroje (voda, energia, potraviny) sa stanú kľúčovou témou nadnárodných rokovaní! Produkcia ropy aj plynu má zostupnú tendenciu! Má sa sústrediť do nestabilných regiónov. Svet asi prejde od ropy k zemnému plynu, uhliu a iným alternatívam. A to aj napriek tomu, že najmä v USA a Karibiku sa našli dosiaľ neobjavené obrovské zdroje ropy a plynu...

Neudržateľný je najmä nedostatočný prístup k zásobám vody i deficit pôdy! Zvlášť sa to týka poľnohospodárstva. Problém sa stupňuje aj vplyvom rozširujúcich sa priemyselných činností a rastu počtu obyvateľstva sveta. Experti označili 21 a potenciálne až 36 štátov, ktoré majú či budú mať problémy nielen s nedostatkom vody, ale aj úrodnej pôdy.

Zmeny klimatických podmienok budú mať ďalší negatívny dopad na nedostatok strategických zdrojov. Mnoho krajín bude bojovať práve s nedostatkom pitnej vody a znižovaním poľnohospodárskej výroby, kde sa očakávajú výrazné straty.

Do roku 2030 sa má zvýšiť počet obyvateľstva na svete o 1,2 mld. ľudí a dopyt po potravinách o 50%! Je to aj v dôsledku rastu blahobytu, vzrastajúcej obľube nižšej strednej triedy o západné stravovacie návyky...

Riešením môžu byť nové technológie, zlomové technologické inovácie a alternatívne zdroje. Tieto môžu napr. vyvinúť alternatívy fosílnych palív, riešenie nedostatku pitnej vody a potravín. *Kľúčové bude tempo technologických inovácií*. Aj napriek politickej a inej podpore bude však prechod na nové zdroje energie (biopalivá, prané uhlie, vodík a pod.) dosť pomalý. *Energetický sektor sa má plne adaptovať na nové zdroje energie za 15 až 25 rokov*. Obnoviteľné zdroje energie (fotovoltaické, slnečné či veterné elektrárne), zdokonaľovanie batériových technológií, rozšírenie hybridných elektrických áut, budovanie domácich vodíkových staníc v garážach, ktoré využívajú na výrobu vodíka elektrinu, môžu značne znížiť náklady, zvýšiť efektívnosť, ziskovosť, resp. konkurencieschopnosť.

Terorizmus, konflikty a šírenie zbraní hromadného ničenia zostanú nedoriešenými kľúčovými problémami!!! Znižovanie terorizmu sa výrazne viaže na *znižovanie nezamestnanosti najmä mladých ľudí* (najmä na Strednom východe a pod.). Hrozbou je aj tvorivé využívanie nových vedecko-technických objavov v tomto smere... Nebezpečenstvom pre celý svet budú „efekty" z nedostatku lásky, nízkej morálky, etiky, resp. rastúca zloba, agresia, túžba po moci, karierizmus, despotizmus, radikalizmus...

Riziko použitia jadrových zbraní sa má zvýšiť. Je to v dôsledku niekoľkých paralelných trendov: rastie počet štátov disponujúcich jadrovými zbraňami, jadrové know-how môžu získať radikálne teroristické skupiny...

Pravdepodobný bude aj trend k rozširovaniu autority a moci vo svete. To môže posilniť vznik ďalších nadnárodných problémov. Má sa posilniť rola nevládnych organizácií (na riešenie rozličných špecifických otázok).

Rozhodujúce faktory na zmenu trajektórií súčasných trendov sú: *zlomové technológie/technologické inovácie*, prisťahovalectvo, zlepšenie zdravotníctva, skvalitňovanie zákonov na podporu *väčšej účasti žien v ekonomike a manažmente, väčšie spojenie ekonómie, etiky a sociálnych dimenzií riadenia* organizácií, podnikov i štátu (dopln. Ľ. J.).

Uvedené trendy teda naznačujú aj značné diskontinuity vo vývoji, v manažmente, nové otrasy, šoky, hrozby... Potrebný je kvalitný preventívny a časový manažment, správne načasovanie fáz ďalších transformácií.

Neisté sú aj dôsledky demografických zmien, ktorým čelí Európa, Japonsko, Rusko atď. Všetko sú to závažné výzvy pre lídrov a manažérov na celom svete!!!

Vyššie spomenuté expertné predpovede vývoja sveta nabádajú „našívať" na mieru modely, metódy, nástroje a techniky/technológie manažmentu. Inými slovami – potrebujeme čo najpresnejšie identifikovať nové paradigmy a podstatné rozdiely takpovediac medzi takpovediac manažmentom prvej generácie (ďalej aj v skratke M 1.0) a manažmentom druhej generácie (M 2.0).

3. Rámcové diferencie medzi manažmentom prvej generácie a manažmentom pripravovanej druhej generácie

Rozdiely medzi „starým" (manažmentu 1.0) a novo sa rodiacim manažmentom (manažment 2.0) možno v podstatných črtách porovnať aj tak, ako to približuje nasledujúca tabuľka (ide samozrejme len o niektoré vybrané konfrontačné „dvojičky").

Rámcové diferencie medzi M 1.0 a M 2.0 Tabuľka

	Manažment 1.0	**MANAŽMENT 2.0**
A	Hierarchická byrokracia, „papierové" kancelárie, slabý manažment času, neplodné schôdzovanie atď.	**Eliminovanie byrokratickej pyramídy – tlakom relevantných inovácií, využívaním IKT najnovších generácií...**
B	Formálne štruktúry, centralizácia riadenia, strohé príkazy zhora, vojenský štýl vedenia ľudí a pod.	**Neformálne štruktúry, decentralizácia manažmentu, participatívny štýl vedenia ľudí.**
C	Vertikálne riadenie a kontrolovanie podriadených zhora nadol. Oslabené modely (vnútro)podnikania.	**Horizontálny manažment aj kontrola, nové (vnútro)podnikateľské modely, akcent na výskum, vývoj, inovácie!!!**
D	Tzv. tvrdé nástroje/prostriedky riadenia. Dôraz na HW, na IQ.	**Tzv. mäkké manažérske prostriedky (SW, imidž). Symbióza: IQ+EQ+MQ!**
E	Krátkodobé a strednodobé horizonty. Tradičné, nekomplexné riadenie. Absencia systémového prístupu.	**Dlhodobé vízie a stratégie, foresight. Procesný, integrovaný, holistický, systémový (celostný) manažment.**
F	Orientácia redukovaného marketingu a obchodovania najmä na lokálne trhy (prevažne bez IKT).	**Globálny marketingový i komerčný záber. Využívanie najnovších IKT, systémov BI (Business Inteligence).**

G	Klasické modely podnikania a komercie (dáta, informácie).	**E-commerce, e-business, knowledge modely riadenia (poznatky, múdrosť).**
H	Bez internetu, resp. web 1. generácie.	**Internet/web 2.0 (sociálne siete atď.)**
I	Manažment = prevažne ekonómia, resp. politická ekonómia, (irelevantné) financovanie.	**Manažment = inter- a multi-disciplinárna veda – aj ekonómia, aj právo, sociológia, psychológia atď.**
J	Postupné odlúčenie etiky, sociológie aj ekológie od ekonómie a riadenia. Znalosti sa považovali najmä za istý vonkajší faktor rozvoja.	**Integrovanie aj etiky, ekológie, sociológie do ekonómie a manažmentu. Znalosti = najmä interný faktor rastu č.1! Dôraz na trvalo udržateľný rozvoj.**
K	Väzby najmä na cudzie (štátne) zdroje v modeli financovania.	**Väčší akcent na vlastné zdroje, bežné aj investičné, viaczdrojové financovanie.**
L	Útlm zodpovednosti vlastníkov, lídrov a manažérov. Orientácia len na práva. Značný deficit kompetentnosti v riadení na všetkých úrovniach.	**Posilnenie zodpovednosti, práv aj povinností vodcov a manažérov. Nárast znalostí, kompetentnosti v manažovaní počnúc trvalou edukáciou, R&D, KPV!**
M	Lojalita zamestnancov k firme sa postupne oslabovala aj pre neúctu k nim (zamestnanec sa začal považovať len za „číslo“...).	**Posilňovanie vzťahov manažéri vs. zamestnanci. Zamestnanec je de facto významný vnútropodnikový zákazník, významne participuje na rozhodovaní!**
N	Orientácia zvlášť na krátkodobé, najmä niektoré ekonomické riziká a hrozby, bez prevencie.	**Včasné holistické, celostné, systémové identifikovanie dlhodobých horieb/rizík a rozvoj preventívneho manažmentu!**
O	Menej kvalitné SWOT analýzy bez preventívneho utlmovania hrozieb. Pohľad zvnútra organizácie navonok.	**Prechod od SWOT k TOWS analýzam, dôslednému preventívnemu utlmovaniu/eliminovaniu rizík. Pohľad zvonku podniku dovnútra!**
P	Dominantný prakticizmus vodcov a manažérov. Nedoceňovanie a nevyužívanie teórie manažmentu a marketingu. Nefunkčný systém výchovy manažérov.	**Integrovať novú vedu omanažmente so skúsenosťami z praxe. Sústavná edukácia (výchova vzdelávanie) manažérov!!! Znalostno-skúsenostný podnikateľský manažment.**
R	Zanedbávanie IKT, internetu a iných hi-tech v riadení, marketingu, komercii, biznise.	**Integrovanie manažmentu 2. generácie s webom 2.0 aj inými vznikajúcimi najnovšími generáciami IKT.**
S	Globálna kríza sa považovala len za ekonomicko-finančnú záležitosť. Chýbali: celostné pohľady, išlo o nekomplexné poznanie a nesystémové riadiace postupy.	**Globálna kríza je aj etický, morálny, sociálny, ekologický, bezpečnostný, politický a iný fenomén či zložitý, komplexný problém. Dominovať majú: tímová múdrosť, zhlukové inovácie atď**
T	Orientácia na rýchle výnosy, tržby a okamžité zisky (za každú cenu). Evidentný egoizmus, sebectvo vodcov a riadiacich pracovníkov!!!	**Postupné dosahovanie relevantných ziskov, resp. rentability celkového disponibilného kapitálu vo väzbách na trvalo udržateľný komplexný rozvoj .**
U	Klasická konkurencia najmä na lokálnych trhoch.	**Super- či hyperkonkurenčné tlaky aj na globálnych trhoch.**

V	Dlhé inovačné cykly, dominancia najmä inkrementálnych inovácií a to zvlášť výrobkov. Slabší manažment zmien, inovácií, výskumu. Rezervy v projektovom manažmente a v podnecovaní ľudí k tvorivosti.	**Krátke inovačné cykly aj výrobkov, služieb, procesov, technológií atď. Dôraz na manažment predvýrobných procesov – výskumu, vývoja, komplexnej prípravy výroby (KPV).** **Vysoká motivácia ľudí ku kreativite!!!**
Z	Ostatné diskrepancie, rozdiely (politické, mocenské, demografické, bezpečnostné, iné deficity) a ich negatívny vplyv na riadenie a ľudí.	**Odklon od bipolárnosti sveta. Presun mocenských centier sveta (aj do Číny, Indie...). Boj o eliminovanie zla, hladu, terorizmu, zbrojárskych tlakov atď.**

4. Záverom

Aj k napísaniu tohto príspevku nás inšpirovali najmä myšlienky hosťujúceho profesora strategického a medzinárodného manažmentu na London Business School, spoluzakladateľa medzinárodnej poradenskej spoločnosti STRATEGOS a riaditeľa konzorcia MANAGEMENT INNOVATION LAB. **Garyho Hamela**, ktorý žije v Severnej Karolíne. Považuje sa za jedného z najuznávanejších súčasných odborníkov v odbore manažmentu na svete. Dokonca denník *Finanacial Times* ho označil za „**inovátora manažmentu, ktorý nemá sebe rovného“**... Postupná tvorba zásadne inovovaného modelu, metód a inštrumentária manažmentu pre budúcnosť si však – aj v SR aj v EÚ – vyžaduje intenzívnu prácu celých tímov renomovaných autorov/expertov.

4. Literatúra

[1] HAMEL G. – BREEN, B.: Budoucnost managementu. Praha, Management Press 2008, 248 s. ISBN 9788072611881.

[2] Global Trends 2025: A Transferred World 2025 (www.dni.gov/nic/NIC_2025_project.html).

[3] JEMALA, Ľ.: Stratégia a systém manažmentu predvýrobných procesov. *Výzva pre lídrov a manažérov učiacich sa organizácií i štátu po roku 2000.* Bratislava, Vydal Ľubomír Jemala 1998, 332 s. ISBN 80-900467-1-1.

[4] JEMALA, Ľ.: Podnikateľský manažment a marketing. Bratislava, Vydavateľstvo STU 2008, 312 s. ISBN 978-80-227-2860-7.

[5] JEMALA, Ľ.: Souřadnice rozvoje managementu. Moderní řízení, XLV., 2010, č. 5, s. 52 – 53.

[6] JEMALA, Ľ.: O 10 až 15 rokov sa očakáva nová kríza aj prevrat storočia v oblasti manažmentu. Zrno, XXI., 2010, č. 16, s. 14 – 17 (úvod do diskusie).

[7] JEMALA, M.: Vstup do metodológie foresightu a príčin možných procesných nedostatkov. Studia commercialia Bratislavensia. Ročník/Volume 3, Číslo/No. 9 (1/2010). ISNN 1337-7493, s. 44 - 59.

[8] JEMALA, M.: Innovative Methodics of Technology Management Assessment and its Criteria. Ekonomika a management organuzací – výzkum, výuka a praxe. Recenzovaný sborník příspěvkú mezinárodní vědecké konference. Brno, 9, září 2010, s. 124 - 132. CD-ROM, ISBN 978-80-210-5273-4.

[9] JIRÁSEK, J. A.: Management budoucnosti (řízení z prvního sledu). Praha, Professional Publishing 2008. ISBN 978-80-86946-82-5.

[10] JIRÁSEK, J. A.: Na obzoru převrat managementu? Moderní řízení, XLV., 2010, č. 2. s. 16 až 18.

[11] KISLINGEROVÁ, E.: Je třeba najít novou strategii. Moderní řízení, XLV., 2010, č. 2, s. 19.

[12] KISLINGEROVÁ, E.: Podnik v čase krize. Praha, Grada Publishing 2009.

[13] Kolektív: Vízia a stratégia rozvoja slovenskej spoločnosti. Úplná verzia. (Podklad na verejnú diskusiu). Bratislava, február 2010, 522 s.

[14] KOPČAJ, A.: Spirálový management *Hledání jádra dynamiky růstu bohatství.* Alfa Publishing, SILMA 90, 2007, 268 s. ISBN 978-80-86851-71-6.

[15] KOŠTURIAK, J.: Manažéri a zmena systému. *Diskusia pokračuje!* Zrno, XXI., 2010, č. 26, s. 14 – 17. *(Diskusia k príspevku [6]).*

[16] MARTINICKÝ, P.: Manažéri a zmena systému (1. časť). Zrno, XXI., 2010, č. 23, s. 11 – 13. *(Diskusia k [6]).*

[17] MARTINICKÝ, P.: Manažéri a zmena systému (2. časť). Zrno, XXI., 2010, č. 24, s. 10 -12. *(Diskusia k [6]).*

[18] SOUČEK, Z. – KOŘENÁ, D.: Předpověď vývoje světa do roku 2025. Moderní řízení, XLV., 2010, č. 4, s. 18 – 22.

[19] ZAJKO, M. – CHODASOVÁ, Z. – JEMALA, Ľ. – MATERÁK, M.: Riadenie malých a stredných podnikov. Bratislava, Vydavateľstvo STU 2010, ... s. ISBN 978-80-227-3344-1.

[20] ZELENÝ, M. – KOŠTURIAK, J.: Do východiskového bodu krízy sa nikdy nemôžeme vrátiť. Zrno, XXI., 2010, č. 9, s. 10 – 14.

[21] www.dni.gov/nic/NIC_2025_project.html).

[22] gh@managmentlab.org.

Adresa autora:

Ľubomír JEMALA, Doc. Ing. PhD.
Oddelenie ekonomiky a manažmentu podnikania
Ústav manažmentu STU
Vazovova 5, 812 43 Bratislava
lubomir.jemala@stuba.sk

Príspevok sa viaže na grantovú úlohu OEMP ÚM STU – VEGA č. 1/0536/10: Inovácie ako strategický základ zvyšovania konkurenčnej schopnosti SR. Smerovanie, meranie a podpora inovačných procesov.

Vliv pracovní neschopnosti na makroekonomickou výkonnost ekonomiky v České republice
Effect of incapacity to work on the macroeconomic performance of the economy in the Czech Republic

Věra Jeřábková, Kristýna Vltavská

Abstract: Incapacity to work is an important indicator not only in the field of the state in population's health but also in the area of economic policy. Labour is one of the main components of production, and if a person is temporarily acknowledged as incapable to work due to medical reasons it influences the person's life as well as the macroeconomic performance of the economy and public finance. The aim of this paper is to analyze the development of incapacity to work due to disease or injury between 2004 and 2009, to find out the most common causes of incapacity to work and estimate the loss in macroeconomic performance of the economy inflicted by incapacity to work.

Abstrakt: Pracovní neschopnost je významným ukazatelem nejen v oblasti zdravotního stavu obyvatelstva, ale i v oblasti hospodářské politiky. Práce je jedním z hlavních výrobních faktorů a v případě, kdy je osoba ze zdravotních důvodů dočasně uznána neschopnou k výkonu svého zaměsnátní, objevuje se zde vliv jak na život daného jedince, na makroekonomickou výkonnost ekonomiky, tak i na veřejné finance. Cílem tohoto příspěvku je analyzovat vývoj pracovní neschopnosti pro nemoc a úraz v letech 2004 až 2009, zjistit nejčastější zdravotní příčiny vzniku pracovní neschopnosti a vyčíslit ztrátu v makroekonomické výkonnosti ekonomiky v souvislosti s pracovní neschopností.

Key words: incapacity to work, macroeconomic performance, macroeconomic loss

Klíčová slova: pracovní neschopnost pro nemoc a úraz, makroekonomická výkonnost ekonomiky

JEL classification: I18, E00.

1. Foreword

Incapacity to work is an important indicator not only in the field the state in population's health but also in the area of national economy. Labour is one of the main components of production and if a person is acknowledged as incapable to work due to medical reasons it influences the person's life as well as the macroeconomic performance of the economy and public finance.

2. Data and Methodology

The observation of the data on incapacity to work has been based mainly on the statistical report Nem Úr 1-02 about the incapacity to work due to disease or injury, processed by the Czech Statistical Office. It is filled in by all economic subjects that employ more than 25 employees/policyholders within the public sector. Members of the Ministry of Defence and the Ministry of Justice (i.e. the Police, Fire brigade, Board of Customs, Prison Service, Intelligence Security Service, Army and Bureau of Foreign Relations and Intelligence) have not been included in the statistics. The focus is on the number of policyholders, the amount of new cases of incapacity to work according to the kind (disease, occupational injury, other injuries) and the length of incapacity to work in calendar days. All these data are divided according to the regions (NUTS3) and sector (CZ-NACE) with a half-year periodicity of

survey. The data of economic subjects with less employees and self-employed workers have been taken from the Czech Social Security Administration.

Besides the statistics of new cases of incapacity to work taken from the data of the Czech Statistical Office, the data on the termination of incapacity to work are processed from the registry of the Czech Social Security Administration via the Institute of Health Information and Statistics of the Czech Republic based on the Decision on Temporary Incapacity to Work document, issued by MDs. This statistics reflects all diseases and injuries causing policyholders at least one-day incapacity to work finished within a given year divided by age, cause of incapacity to work (MKN-10 diagnosis) and occupation classification KZAM. Therefore, injuries or diseases resulting in incapacity to work overlapping to the following year or those cases where the Decision on Temporary Incapacity to Work was not issued are not included.

Differences in numbers of cases of incapacity to work between the aforementioned sources reflect the differences in methodology of data collection. This paper will only use the data of the Czech Statistical Office with causes of incapacity to work taken from the Institute of Health Information and Statistics of the Czech Republic.

Data on GDP in current prices, average number of registered employment, average percentage of incapacity to work and employee compensations will be crucial for estimation of the loss in macroeconomic performance in the economy. Average percentage of incapacity to work is a total indicator with respect to both new cases of incapacity to work and the length of the cases in relation to the number of health policyholders and calendar fund. Compensations for employees incapacitated to work can be calculated based on the following relation: (1):

$$NZ_{PN} = \left(\frac{NZ}{E} \right) \cdot PN \ , \tag{1}$$

where
NZPN – compensations for employees incapacitated to work,
NZ – employee compensations,
E – average number of registered employees,
PN – average number of cases of incapacity to work absent from the workplace.

3. Development in incapacity to work due to disease or injury between 2004 and 2009

Incapacity to work relates closely to the area of health insurance because in case of losing the source of income due to a so-called short-term social accident (i.e. a temporary incapacity to work due to disease or injury, care for a family member, pregnancy or motherhood, care for a child), the person is sustained by benefits emerging from this insurance. Among the policyholders of health insurance belong – obligatorily – all employees and – optionally – self-employed workers. Health insurance is based on a solidarity system and potential changes in legislation necessarily change the development of incapacity to work. The level of incapacity to work is not given only by objective causes for the emergence of a disease or injury but often depends on wider circumstances, such as changes in legislation in health insurance or the labour market situation.

Incapacity to work is in most cases starts due to a disease. The ratio of incapacity to work due to a disease on the overall number of newly reported cases of incapacity to work due to disease or injury reaches an average of 92 % in the given period, i.e. between 2004 and 2009. The overall number of newly reported cases of incapacity to work per 100

policyholders ranged from 33.89 to 58.19 with 2005 being the year with the highest number and 2009 the lowest. Figure 1 indicates that a permanent fall has been registered since 2005 and the average drop in the given period was 11 %. The largest year-to-year drop was registered in 2009 when the new cases per 100 policyholders fell by 30 % compared to 2008 and the number of newly registered cases of incapacity to work dropped by 35 %. The reason for such a change lies mainly in accepting the new law No.187/2006 Coll. on health insurance that came into force on 1st January 2009. The main amendment changed the way to determining the extent of health insurance benefits, which also influenced the registration methodology, drawing the benefits, data collection and data processing in the Czech Social Security Administration. However, the amount of days on a sick-leave, which ranged from 33 to 45 days in the given period, grew in the period with an annual average of 5 % (app. 2 days a year in absolute numbers).

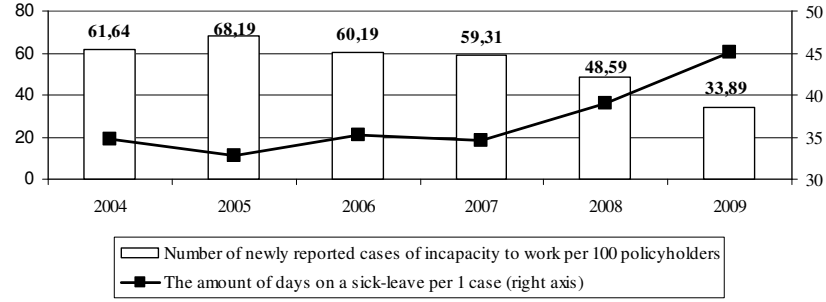


Fig. 1: Number of newly reported cases of incapacity to work per 100 policyholders and the amount of days on a sicl-leave per 1 case (right axis) in the period between 2004 and2009
Source: Czech Statistical Office

In terms of regions, the lowest number of cases of incapacity to work per 100 policyholders between 2004 and 2008 was recorded in Prague (42.21 to 59.38 cases). In 2009 this spot was taken by Olomoucký region where only 31.65 new in 100 policyholders were registered. On the other hand, the highest numbers of new cases in the given period were recorded in Liberecký and Plzeňský region (table1).

The highest number of incapacity to work between 2004 and 2009 were reported in manufacturing industry where the average number of new cases of incapacity to work per 100 policyholders reached 76.36.

In terms of the causes of incapacity to work, the most common reasons were diseases of the respiratory system (fig.2) with the ratio on the overall numbers constituted 33 - 47 % in the given period. Other common causes were diseases of the musculosketal system and connective tissue, which formed in average a fifth in the overall incapacity to work while their ratio was rather stable between the years 2004 and 2009. Injury, poisonings and other consequences of external causes (12 % on average), diseases of the digestive system (7 % on average), diseases of the genitourinary system (4 % on average) and diseases of the circulatory system (3 % on average) were other major causes. Altogether these six groups of diagnoses built on average 84 % of the overall numbers of incapacity to work in the period between 2004 and 2009. As figure 2 shows, the ratio is stable within all groups mentioned above apart from the diseases of the respiratory system where the ratio differs significantly from year to year. The reasons for the prominent changes can be found in seasonal influenzas

and epidemics of influenza, legislative changes that affect the interest of the sick to visit the doctor because of the disease and get incapacitated to work.

Table 1: Number of newly reported cases of incapacity to work per 100 in Czech regions in the period between 2004 and 2009

Region	2004	2005	2006	2007	2008	2009
Praha	**53.38**	**59.18**	**52.19**	**51.10**	**42.21**	32.70
Středočeský	60.17	67.30	60.03	58.99	49.50	36.28
Jihočeský	65.67	72.96	63.57	63.07	50.78	34.87
Plzeňský	67.51	76.84	65.88	64.57	54.44	37.34
Karlovarský	65.83	72.57	63.61	63.07	53.13	34.91
Ústecký	59.99	65.38	58.11	58.54	48.84	32.71
Liberecký	68.31	74.34	66.44	64.98	54.04	36.72
Královéhradecký	65.16	72.27	62.58	62.81	48.82	33.61
Pardubický	64.53	70.48	63.07	61.44	49.67	33.61
Vysočina	65.88	74.34	64.43	64.07	51.59	33.73
Jihomoravský	62.86	68.74	61.76	59.77	47.96	33.36
Olomoucký	65.15	68.56	61.03	59.39	48.61	**31.65**
Zlínský	64.79	70.67	62.87	61.27	50.30	33.44
Moravskoslezský	61.63	70.62	62.50	63.88	52.14	33.84
Česká republika	61.64	68.19	60.19	59.31	48.59	33.89

Source: Czech Statistical Office

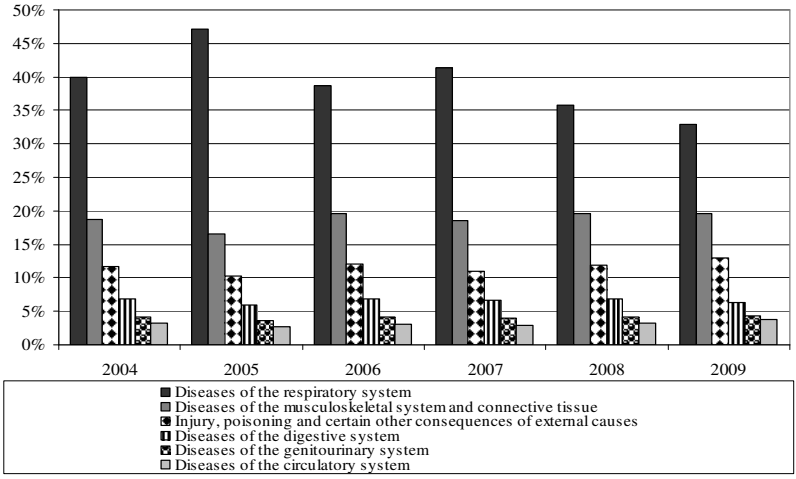


Fig. 2: Development of the most common causes of incapacity to work on the total number of incapacities to work in the period between 2004 and 2009 (%)

Source: ÚZIS

4. The loss in macroeconomic performance of economy

Resulting from incapacity to work, hundreds of thousand employees are absent from their workplace daily, which to a certain extent affects both the microeconomic and the macroeconomic area of the national economy. Social benefits, health care expenditures for policyholders incapacitated to work, lower insurance and tax collection (neither insurance nor tax is paid from the salary compensations) impact the public finance. The following text focuses only on the enumeration of the loss in macroeconomic performance of economy due to incapacity to work.

This estimate uses the data on the average number of employment, average percentage of incapacity to work and employee compensation and gross domestic product, i.e. the overall financial value of goods, estates and services produced in a given period of time, which economy uses for determining the performance of an economy.

The GDP in current prices reached CZK 3,625,865 million in 2009. The average number of registered employment (physical entities) in the Czech Republic amounted to 3,960,500 people and the average percentage of incapacity to work was 4.184 %. On average, 165,707 employees were absent from the workplace in 2009. In the same year, an average worker produced what equals CZK 955,484 in GDP. The overall GDP might have been higher by CZK 158,331 if all employees worked. Overall employee compensations for those incapacitated to work, decreasing the hypothetical loss in GDP, were CZK 67,259 million and were calculated as shown in (1).

A rough estimate of the loss in macroeconomic performance in national economy may be calculated by subtracting the overall compensations to the incapacitated to work from the hypothetical GDP if all employees worked. In 2009, this loss was estimated to be CZK 91,072 million and one percent of incapacity to work constituted a loss of CZK 21,767 million.

Rough estimates of macroeconomic loss for years 2004 to 2008 can be calculated analogically. Table 2 clearly shows that between 2004 and 2007 such loss grew annually by CZK 6,860 million on average. The year 2008 started the decrease, however. In 2009, the year-to-year drop reached 22 %, mainly a consequence of the decrease of average percentage of incapacity to work due to disease or injury following the changes in legislation.

Table 2: Rough estimates of the loss in macroeconomic performance in economy in the period between 2004 and 2009

	2004	2005	2006	2007	2008	2009
GDP in current prices (mil. CZK)	2 814 762	2 983 862	3 222 369	3 535 460	3 688 997	3 625 865
Average number of registered employment (in thousands)	3968,3	4035,6	4067	4137	4188,1	3960,5
Average percentage of incapacity to work	5,857%	6,126%	5,814%	5,619%	5,184%	4,184%
Hypothetical increase of GDP (mil. CZK)	175 117	194 720	198 913	210 485	201 693	158 331
employee compensations (mil. CZK)	70 413	78 699	80 595	85 201	84 679	67 259
The loss in macroeconomic performance in the economy (mil. CZK)	**104 704**	**116 021**	**118 319**	**125 283**	**117 014**	**91 072**

Source: Czech Statistical office, computations of authors

5. Conclusion

Incapacity to work is an important indicator not only in the field of the state in population's health but also in the area of economic policy because it influences the person's life as well as the macroeconomic performance of the economy and public finance.

The development of incapacity to work is mainly connected to changes in legislation regarding health insurance which determines the extent of the benefits, to changes on the labour market and occurrence of individual diseases. The main impact on incapacity to work is inflicted by the diseases of the respiratory system (from the MKN-10 disease groupings). The ratio of these diseases varies significantly mainly due to seasonal influenzas and epidemics.

This paper focused only on estimating the loss in macroeconomic performance of the national economy. The estimate drew data from the GDP in current prices, average number of registered employment, average percentage of incapacity to work and employee compensations. Rough estimates of loss in macroeconomic performance of national economy between 2004 and 2009 ranged from CZK 91,072 million (year 2009) to CZK 125,283 million (year 2007).

The reason for the year 2009 to register the highest annual decrease both in the number of newly registered cases of incapacity to work and in the estimated loss in macroeconomic performance of national economy was accepting of the law No.187/2006 Coll. on health insurance. However, a new trend can be traced in the data. While the number of newly registered cases of incapacity to work started to decrease, the number of days on a sick-leave is grown.

6. References

[1] CZECH STATISTICAL OFFICE. Pracovní neschopnost pro nemoc a úraz. Data za roky 2004 až 2009.

[2] CZECH STATISTICAL OFFICE. Nem Úr 1-02 Výkaz o pracovní neschopnosti pro nemoc a úraz. Dostupné z: http://www.czso.cz/csu/klasifik.nsf/i/nem_ur_1_02_psz_2010.

[3] CZECH STATISTICAL OFFICE. Vývoj pracovní neschopnosti pro nemoc a úraz v letech 1990-2003: http://www.czso.cz/csu/2005edicniplan.nsf/p/1127-05

[4] DEMOGRAFIE. KLESL, ARNOŠT. Pracovní neschopnost – faktor omezující produktivitu práce: http://www.demografie.info/?cz_detail_clanku&artclID=512.

[5] MINISTRY OF LABOUR AND SOCIAL AFFAIRS. Health insurance: http://www.mpsv.cz/cs/7.

[6] ÚSTAV ZDRAVOTNICKÝCH INFORMACÍ A STATISTIKY. Ukončené případy pracovní neschopnosti pro nemoc a úraz v České republice v roce 2009.

[7] LAW No.187/2006 Coll. on health insurance.

[8] WORLD HEALTH ORGANIZATION. INTERNATIONAL STATISTICAL CLASSIFICATION OF DISEASES AND RELATED HEALTH PROBLEMS, 10TH REVISION.

Contacts:

Věra Jeřábková, Ing.
KEST VŠE Praha
nám. W. Churchilla 4
130 67 Praha
vera.jerabkova@vse.cz

Kristýna Vltavská, Ing.
KEST VŠE Praha
nám. W. Churchilla 4
130 67 Praha
kristyna.vltavska@vse.cz

This paper has been prepared under the support of the project of the University of Economics, Prague - Internal Grant Agency, project No. 30/2010 "Multifactor Productivity Estimates".

Analýza extrémnych hodnôt v poistení motorových vozidiel metódou POT
Extreme value analysis in car insurance by POT method

Matej Juhás, Valéria Skřivánková

Abstract: This paper deals with analysis of extremal claims in car insurance by peaks over threshold method (POT). Our aim is to give tail estimation of claim distribution using some graphical and analytical methods.

Key words: Generalized Pareto distribution , parameter estimation, peaks over threshold

Kľúčové slová: Zovšeobecné Paretovo rozdelenie, odhad parametrov, excedenty ponad prah

JEL classification: C13, C16, G22

1. Úvod

Výskyt extrémne veľkých poistných plnení v posledných rokoch núti poisťovne pristúpiť k zmenám v hodnotení a riadení ich rizík a vo väčšej miere využívať pokročilejšie štatistické metódy, včítane metód teórie extrémnych hodnôt. Extrémne poistné plnenia môžu spôsobiť insolventnosť poisťovne, preto je veľmi dôležité poznať ich rozdelenie a vedieť predpovedať ich očakávanú hodnotu. Štatistická analýza extrémnych hodnôt vyžaduje špeciálne grafické a analytické metódy, ktoré sú rozpracované napr. v monografiách Embrechts at al. (1997), Beirlant at al. (2004) a Reiss, Thomas (2007).

Rozdelenie extrémnych hodnôt sa vyznačuje tým, že pravý koniec (chvost) jeho hustoty konverguje k nule oveľa pomalšie ako hustoty exponenciálneho typu (napr. normálne, exponenciálne). Rozdelenie má teda „tučný chvost". Modelovať všetky dáta jediným rozdelením v prípade prítomnosti extrémnych hodnôt je dosť obtiažné, preto je dôraz kladený na odhad rozdelenia chvostovej časti.

Postup pri výbere vhodného rozdelenia spočíva v troch krokoch: 1) na základe grafických metód navrhnúť vhodný typ rozdelenia chvosta 2) odhadnúť parametre zvoleného rozdelenia 3) overiť zhodu vybraného rozdelenia s empirickým rozdelením testom.

V ďalšej časti uvedieme základné pojmy a budú prezentované grafické a analytické metódy, potrebné na štatistickú analýzu reálnych dát v časti 3. V závere rozdiskutujeme výhody a nevýhody použitých metód.

2. Základné pojmy a metódy

V tomto príspevku využijeme z teórie extrémnych hodnôt metódu POT založenú na zovšeobecnenom Paretovom rozdelení (GPD), na popis chvosta rozdelenia poistných dát. Predpokladajme, že poistné dáta reprezentujú realizáciu x náhodnej veličiny X ponad dostatočne vysoký prah u. Nech $F(x)$ je distribučná funkcia náhodnej veličiny X a $F_u(x)$ predstavuje podmienenú distribučnú funkciu excedentov X ponad prah u. Potom platí

$$F_u(x) = P(X - u \le x \mid X > u) = \frac{F(x+u) - F(u)}{1 - F(u)}, \text{ pre } x > 0. \tag{1}$$

Parametrické modelovanie distribučnej funkcie $F_u(x)$ excedentov ponad prah u je založené na limitnej vete (pozri [3]), ktorú dokázali Balkema, de Haan a Pickands. Podľa tejto vety pre dostatočne veľké u ($u \to \infty$) distribučná funkcia excedentov $F_u(x)$ konverguje ku distribučnej funkcii zovšeobecneného Paretovho rozdelenia $G_{\xi\beta}$ ($F_u(x) \approx G_{\xi\beta}(x)$) definovanej vzťahom

$$G_{\xi,\beta}(x)=\begin{cases}1-\left(1+\xi\dfrac{x}{\beta}\right)^{-1/\xi} & \text{ak } \xi\neq 0,\\ 1-exp(-x/\beta) & \text{ak } \xi=0,\end{cases} \tag{2}$$

pričom $x\in\langle 0,\infty)$ ak $\xi\geq 0$ a $x\in\langle 0,-\beta/\xi\rangle$ pre $\xi<0$.

Pri metóde POT je veľmi dôležité vhodne zvoliť prah u. Ak je u príliš veľké, tak máme príliš málo excedentov a dôsledkom toho je veľký rozptyl odhadov. Pre u príliš malé sa odhady stávajú vychýlenými. V ďalšom uvedieme niektoré metódy vhodné na určenie prahu.

Prvovu z možností ako vhodne zvoliť prah u je využitie linearity *funkcie priemerných excessov* $e(u)=E(X-u|X>u)$ pre GPD. Funkcia $e(u)$ náhodnej veličiny X s distribučnou funkciou $G_{\xi,\beta}$ je tvaru

$$e(u)=\int_u^\infty \frac{1-F(y)}{1-F(u)}dy=(\beta+\xi u)^{1/\xi}\int_u^\infty(\beta+\xi y)^{-1/\xi}dy=\frac{\beta+\xi u}{1-\xi}, \tag{3}$$

pre $\beta+\xi u>0$, $\xi<1$, a teda je zrejmé, že $e(u)$ je lineárnou funkciou u. Mali by sme preto zvoliť taký prah $u>0$, pre ktorý je funkcia priemerných excessov empirických dát približne lineárna. Ak dáta $x_1, x_2,...x_n$ zoradené od najmenšej po najväčšiu hodnotu označíme $x_{(1)}, x_{(2)}, ..., x_{(n)}$ potom *výberový tvar* funkcie priemerných excessov je

$$e_n(u)=\frac{\sum_{i\leq n}(x_i-u)I(x_i>u)}{\sum_{i\leq n}I(x_i>u)},\ \text{pre } x_{(1)}<u<x_{(n)}. \tag{4}$$

Inú vhodnú metódu na voľbu prahu poskytuje stabilita Hillovho grafu. Nech $X_{1,n}>X_{2,n}>\ldots>X_{n,n}$ je usporiadaný náhodný výber kladných, nezávislých, rovnako rozdelených náhodných veličín. Potom *Hillov odhad parametra* $\xi>0$ rozdelenia $G_{\xi,\beta}$, založený na k najväčších pozorovaniach, je definovaný vzťahom (pozri napr. [1], [2])

$$\xi_{k,n}=\frac{1}{k}\sum_{i\leq k}ln\,X_{i,n}-ln\,X_{k,n}. \tag{5}$$

Hillov graf je daný bodmi $[k,\ \xi_{k,n}]$, $k=1,2,...,n-1$ a vykazuje stabilitu v takej oblasti, kde rastúcim k ostáva odhad $\xi_{k,n}$ relatívne nezmenený. Vhodné je zvoliť prah u práve z takej stabilnej oblasti.

Ďalšia z vhodných metód voľby prahu je založená na minimalizácii výrazu

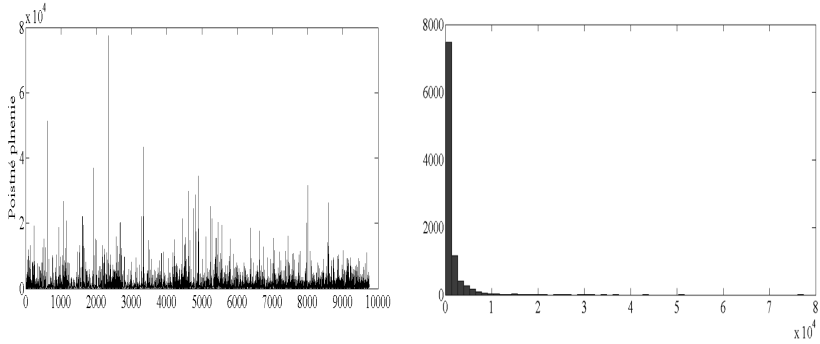
$$\frac{1}{k}\sum_{i\leq k}i^c\left|\xi_{i,n}-med(\xi_{1,n},...,\xi_{k,n})\right|, \tag{6}$$

pričom $0\leq c<1/2$ a $med(\xi_{1,n},\xi_{2,n},\ldots,\xi_{k,n})$ vyjadruje medián odhadov $\xi_{1,n},\xi_{2,n},\ldots,\xi_{k,n}$, vzhľadom na uvažovaný počet extrémov k (podrobnejšie pozri [3]). Je nutné povedať, že takáto metóda hľadania prahu je vhodná len pre výber nie príliš veľkého rozsahu, pretože je časovo dosť náročná. Metódy vhodné na voľbu prahu možno nájsť aj v [1].

3. Analýza dát z havarijného poistenia motorových vozidiel metódou POT

V tejto časti sa budeme zaoberať štatistickou analýzou dát havarijného poistenia motorových vozidiel zo slovenského poistného trhu (poisťovňa, ktorá poskytla dáta, si neželá byť menovaná). Naším hlavným cieľom je odhadnúť parametre rozdelenia chvosta týchto dát. K dispozícii máme súbor dát pozostávajúci z 9748 poistných plnení v eurách v rokoch 1998 až 2008. Analýzu dát sme uskutočnili využitím softwaru R (balík extremes), Matlab (balík evim) a Xtremes, ktorý je prílohou k [3].

Na Obrázku 1 vľavo je znázornený časový rad, ktorý nám umožňuje lokalizovať a určiť výšku extrémnych poistných plnení. Histogram vpravo svedčí tiež o výskyte extrémnych hodnôt a umožňuje nám identifikovať tvar rozdelenia poistných plnení.

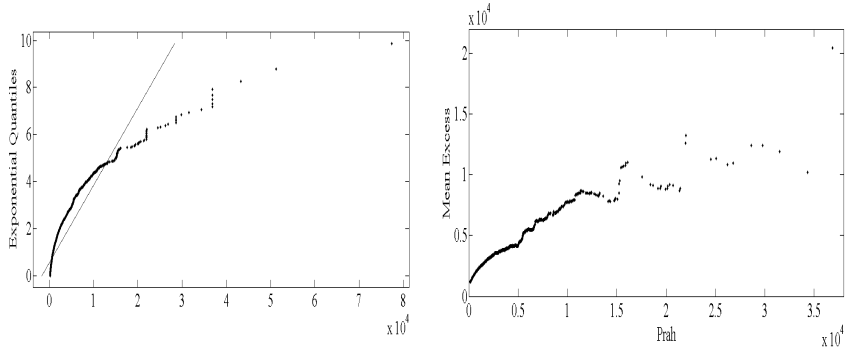


Obrázok 1: Výšky poistných plnení v eurách a odpovedajúci histogram.

Tabuľka7: Základné charakteristiky dát

Počet	Min	Max	Priemer	Medián	Rozptyl	Šikmosť	Špicatosť
9748	150.2	77384.5	1348.458	571.015	6831693	8.5	134.997

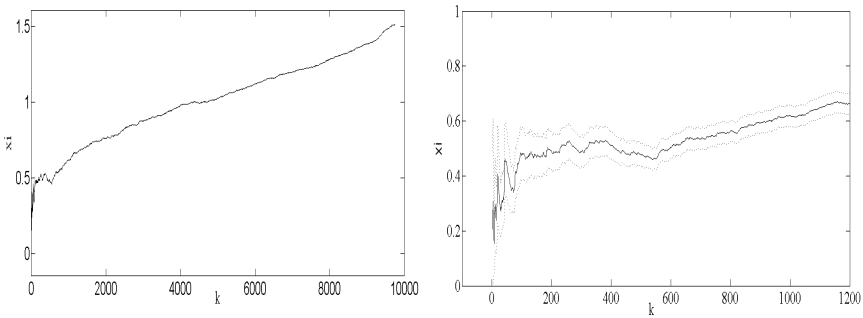
Z tabuľky 1 je vidieť, že špicatosť rozdelenia uvažovaných dát podstatne prevyšuje špicatosť normálneho rozdelenia, ktorá je 3. To znamená, že rozdelenie dát patrí medzi rozdelenia s tučnými chvostmi. V tomto tvrdení nás upevňuje aj to, že QQ graf dát oproti kvantilom štandardného exponenciálneho rozdelenia je konkávny (Obrázok 2 vľavo).



Obrázok 2: Exponenciálny QQ graf (vľavo) a funkcia priemerných excessov (vpravo).

Prvým krokom v analýze poistných dát bol odhad rozdelenia celej vzorky dát. Zistili sme, že dáta je možné modelovať trojparametrickým log-normálnym rozdelením, avšak na popis chvosta rozdelenia týchto dát nie je až také vhodné. Preto sa v ďalšom pokúsime odhadnúť parametre rozdelenia chvosta dát. Na základe Balkema, de Haan a Pickandsovej vety určíme neznáme parametre GP rozdelenia s distribučnou funkciou (2). Najprv si potrebujeme vhodne zvoliť prah *u*. Využijeme pritom graf funkcie priemerných excessov

(zostrojený podľa (3) a (4)) a graf Hillovho odhadu (5) ako funkcie počtu extrémov *k*. Potrebujeme vybrať také *u*, pre ktoré je funkcia priemerných excessov približne lineárna alebo taký počet extrémov (excedentov), pre ktorý je Hillov odhad stabilný.



Obrázok 3: Graf Hillovho odhadu (vľavo), výrez z grafu Hillovho odhadu a interval spoľahlivosti (vpravo)

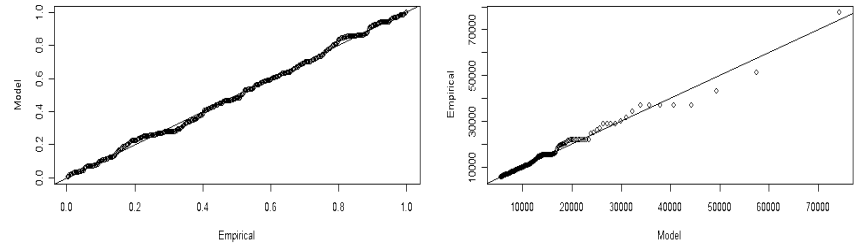
Linearita z Obrázka 2 (vpravo) a tak isto ani stabilita Hillovho odhadu (Obrázok 3) nie je až taká zrejmá a je silne subjektívna. Po detailnom skúmaní obrázkov 2 a 3 považujeme za vhodné zvoliť počet extrémov k = 380 čomu zodpovedá hodnota u = 5520.51. Ak na odhad prahu použijeme metódu založenú na minimalizácii vzťahu (6), tak pri voľbe c blízke nule dostávame hodnotu k = 634, čomu zodpovedá prah u = 4331.71.

Pre tieto hodnoty prahov a odpovedajúci počet extrémov odhadneme parametre GP rozdelenia metódou maximálnej vierohodnosti. Potrebnú hustotu pravdepodobnosti získame deriváciou vzťahu (2) pre $\xi > 0$.

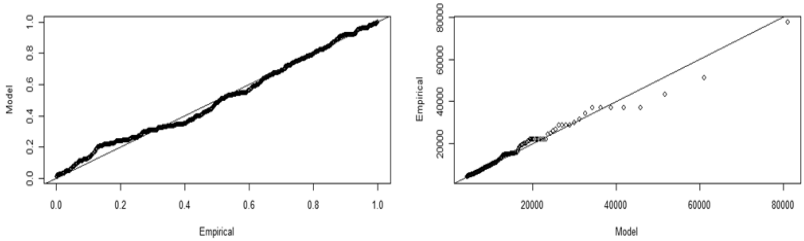
Tabuľka2: Odhady parametrov GP rozdelenia

Počet extrémov	β	ξ	**Hillov odhad**
k =380	3501.222	0.345	0.529
k =634	2553.209	0.396	0.518

Na nasledujúcich obrázkoch uvedieme pravdepodobnostné a kvantilové grafy pre oba počty extrémov a hustotu GP rozdelenia s odhadnutými parametrami uvedenými v tabuľke 2.

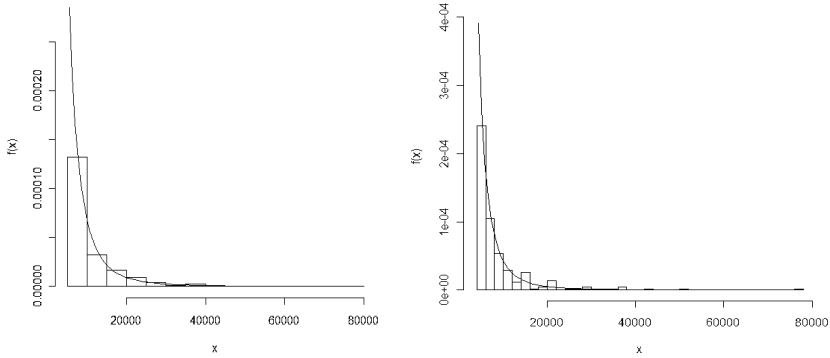


Obrázok 4: PP a QQ graf empirického rozdelenia oproti teoretickému pre hodnotu k =380



Obrázok 5: PP a QQ graf empirického rozdelenia oproti teoretickému pre hodnotu k =634

Obrázok 6: Histogram a hustota pravdepodobnosti s odhadnutými parametrami (vľavo



k=380, vpravo k=634)

Na záver otestujeme zhodu teoretického rozdelenia s empirickým, na čo použijeme Kolmogorovov - Smirnovov (KS) test na hladinách významnosti $\alpha = 0.05$, $\alpha = 0.01$.

Tabuľka3: Kolmogorovov - Smirnovov test

Veľkosť výberu	380		
Štatistika	0.03996		
p-Hodnota	0.56502		
α		0.05	0.01
Kritická hodnota		0.06966	0.08357
Zamietnutie nulovej hypotézy		Nie	Nie
Veľkosť výberu	634		
Štatistika	0.06319		
p-Hodnota	0.0121		
α		0.05	0.01
Kritická hodnota		0.05393	0.0647
Zamietnutie nulovej hypotézy		Áno	Nie

Na základe PP grafov, QQ grafov a výsledkov KS-testu je voľba počtu extrémov k = 380 lepšia ako k = 634. Preto je vhodné modelovať chvost rozdelenia uvažovaných poistných plnení GP rozdelením s parametrami ξ = 0.345, β = 3501.222. Pomocou takto odhadnutého rozdelenia môžeme určiť maximálnu výšku poistného plnenia, ktoré nebude prekročené s pravdepodobnosťou napríklad 0.995. Pre naše dáta je to $x_{0.995}$= 52984.45 eur.

4. Záver

Metódy, ktoré sme použili pri štatistickej analýze reálnych dát, majú svoje výhody aj nevýhody. Grafické metódy sú jednoduché, poskytnú prvotné informácie o výskyte a rozdelení extrémnych hodnôt, majú však aj svoje nedostatky. Hľadanie vhodnej úrovne u nemusí viesť k jednoznačnému výberu rozdelenia. Otestovanie zhody empirických a teoretických výsledkov je preto nevyhnutné. Dáta majú tendenciu sa zhlukovať, podmienka nezávislosti a identického rozdelenia nemusí byť splnená, volatilita sa časom môže meniť. Tým dochádza ku skresleniu odhadov a prognóz.

5. Literatúra

[1] BEIRLANT, J. at al. (2004). Statistics of Extremes: Theory and applications. Wiley, New York.

[2] EMRECHTS, P. at al. (1997). Modelling extremal events for insurance and finance. Springer, New York.

[3] REISS D.-R., THOMAS M. (2007). Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. Birkhäuser Verlag, Basel.

Adresa autorov:

Matej Juhás, Mgr
Ústav matematických vied, PF UPJŠ
Jesenná 5
040 01 Košice
matej.juhas@student.upjs.sk

Valéria Skřivánková, Doc. RNDr. CSc
Ústav matematických vied, PF UPJŠ
Jesenná 5
040 01 Košice
valeria.skrivankova@upjs.sk

Tento príspevok vznikol s podporou grantu VEGA č. 1/0131/09 a VEGA č. 1/0325/10.

Reliability of the data sources on international migration
Věrohodnost údajů o mezinárodní migraci

Eva Kačerová

Abstract: Difficulty in monitoring of migration processes is one of the most important problems connected with research on this phenomenon in various countries. The definition of migrant is not the same in selected countries.

This article came into being within the framework of the institutional support of the long-term conceptual development in science and research on the Faculty of informatics and statistics on the University of Economics, Prague.

Abstrakt: Obtíže při sledování migračních toků jsou jedním z důležitých problémů se studiem migrace spojených. Potíže přinášejí různé definice imigrantů a emigrantů v různých zemích. Nejednotné jsou i údaje o počtech vystěhovalých z pohledu země, kterou migrant opouští, a přistěhovalých z pohledu přijímající země.

Tento článek vznikl s institucionální podporou dlouhodobého konceptuálního rozvoje vědy a výzkumu na Fakultě informatiky a statistiky Vysoké školy ekonomické v Praze.

Key words: Migration, immigration, emigration, bilateral migration flow Central Europe.

Klíčová slova: Migrace, imigrace, emigrace, vzájemné migrační toky ve střední Evropě.

JEL classification: F22, O15.

1. Introduction

In this part we look at available data as seen by the end user. We investigate immigration and emigration data by organising them in a way that allows us to compare data reported by sending and receiving countries and to evaluate international comparability of data provided by individual countries. We analyse two types of information: the double entry matrix containing the flows between selected country (Czech Republic, Slovak Republic, Poland and Hungary) and time series of flows between selected pairs of these countries.

2. Reliability of the data sources on international migration

In order to illustrate the problems with data on international migration flows we have constructed a double entry matrix for the year 2003 and for the year 2005 (tables 1 and 2). The idea of double entry migration matrix is to present the data on immigration, reported by the receiving countries, and those on emigration, reported by the sending countries, in one table. The cells in tables 5 and 6 representing migration from country A to country B contain two entries: the upper one includes immigration (I) form country A reported by country B and the lower one includes emigration (E) to country B reported by country A. For a better understanding the data in a double entry matrix we have calculate I/E ratio and I – E differences, where I and E are the flows reported by the receiving and by the sending country. The figures reported by the receiving country are often several times higher than those reported by the sending country. Large I/E ratio have been observed for flows from Slovakia to Czech Republic (I/E = 54 in 2003, 14 in 2005) and from Poland to Czech Republic (I/E = 36 in 2003, 25 in 2005).

The general believe is that immigration data are better than those concerning emigration. The flow from Slovakia to the Czech Republic in 2003 was according to Czech Republic 24 385 people; the value reported by Slovakia was only 448. The flow from Czech

Republic to Slovakia was 18 262 according the Czech data source and 650 according to Slovakia data source. So, both countries had a positive net migration. The flow from Slovakia to the Czech Republic in 2005 was according to Slovakia 734 people; the value reported by Czech Republic was 10 133. The flow from Slovakia to the Czech Republic was 1 144 according the Czech data source and 1 950 according to Slovakia data source. So, both countries had a positive net migration.

Identifying and counting expatriates is not without difficulties and different methods may produce different estimates. There are three main types of estimates, each of them with it advantages and shortcomings: emigration survey in origin countries and compilation of statistics from receiving countries and population census.

Table 1: Migration flows between selected countries according to receiving (I) and sending (E) countries in 2003

Sending country		Receiving country			
		Czech Republic	Hungary	Poland	Slovak Republic
Czech Republic	I	–	…	46	650
	E	–	35	1 040	18 262
Hungary	I	58	–	20	25
	E	…	–	…	…
Poland	I	1 653	…	–	36
	E	46	6	–	10
Slovak Republic	I	24 385	…	19	–
	E	448	18	10	–

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

Table 2: Migration flows between selected countries according to receiving (I) and sending (E) countries in 2005

Sending country		Receiving country			
		Czech Republic	Hungary	Poland	Slovak Republic
Czech Republic	I	–	…	60	1 144
	E	–	4	138	1 935
Hungary	I	28	–	21	248
	E	…	–	…	…
Poland	I	1 246	…	–	311
	E	49	13	–	5
Slovak Republic	I	10 133	…	31	–
	E	734	28	6	–

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

Other interesting observations can be made by looking at the figures presenting the evolution of the flows between pairs of countries over time reported by each of both countries. Such graphs are very helpful when trying to understand international migration trends and prepare a forecast (Figure 1-6). Dates for Hungary are not available.

The direct comparison of flows between Poland and Slovakia reported by the sending and receiving countries reveals important feature of the statistics based on the concept of permanent place of residence, namely the underestimation of emigration flows. The data reported by the receiving country are higher both for flows from Poland to Slovakia and from Slovakia to Poland.

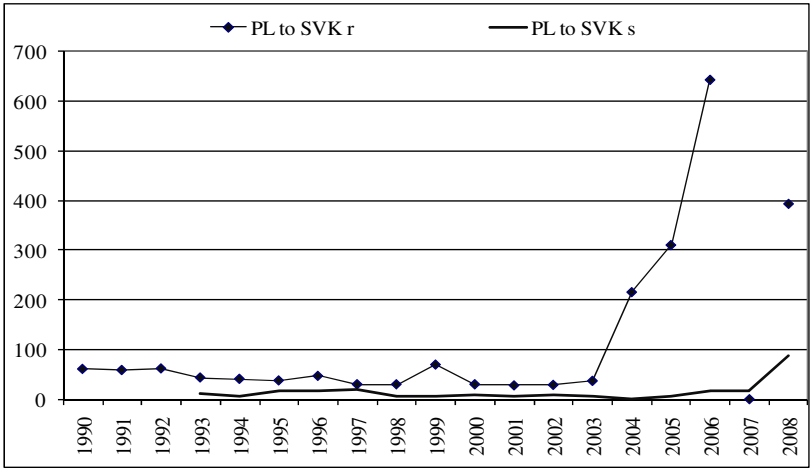


Figure 1: Migration between Poland and Slovak Republic

r – data according to the receiving countries s – data according to the sending countries

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

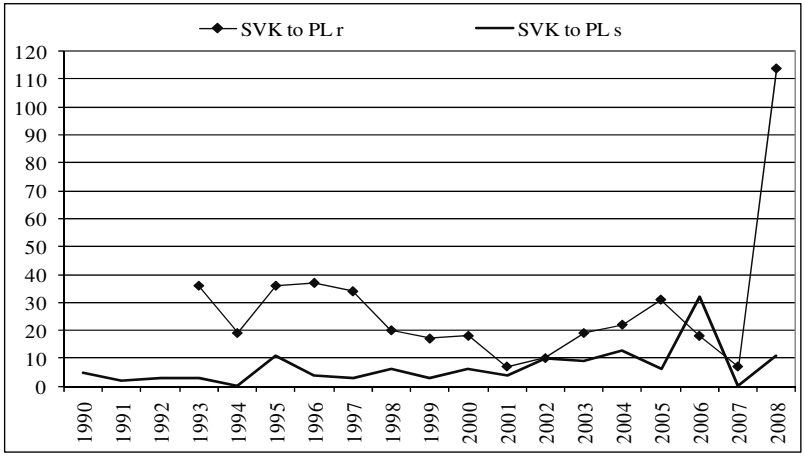


Figure 2: Migration between Poland and Slovak Republic

r – data according to the receiving countries s – data according to the sending countries

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

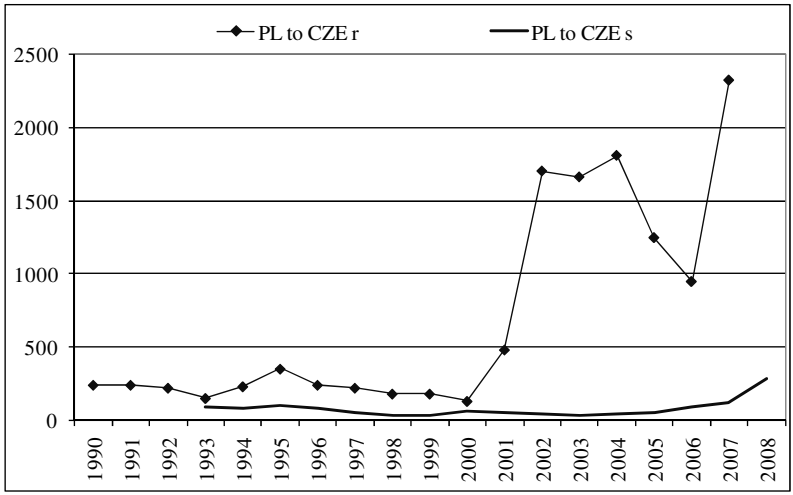


Figure 3: Migration between Poland and Czech Republic

r – data according to the receiving countries s – data according to the sending countries

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

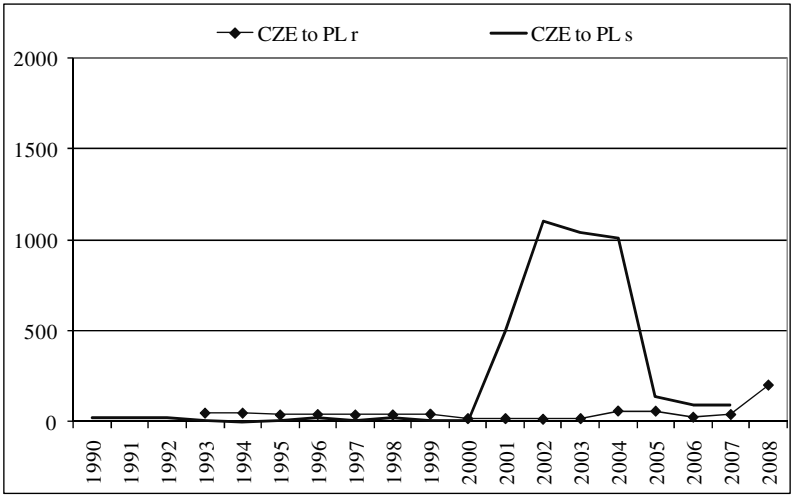


Figure 4: Migration between Poland and Czech Republic

r – data according to the receiving countries s – data according to the sending countries

Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

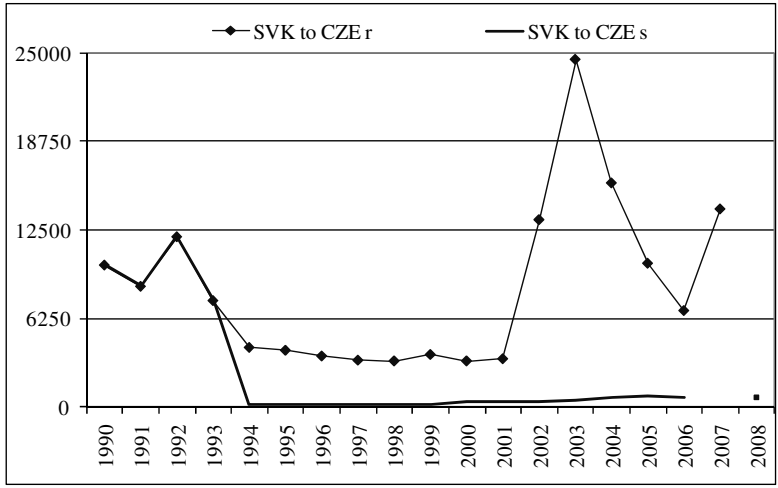


Figure 5: Migration between Slovak Republic and Czech Republic
r – data according to the receiving countries s – data according to the sending countries
Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

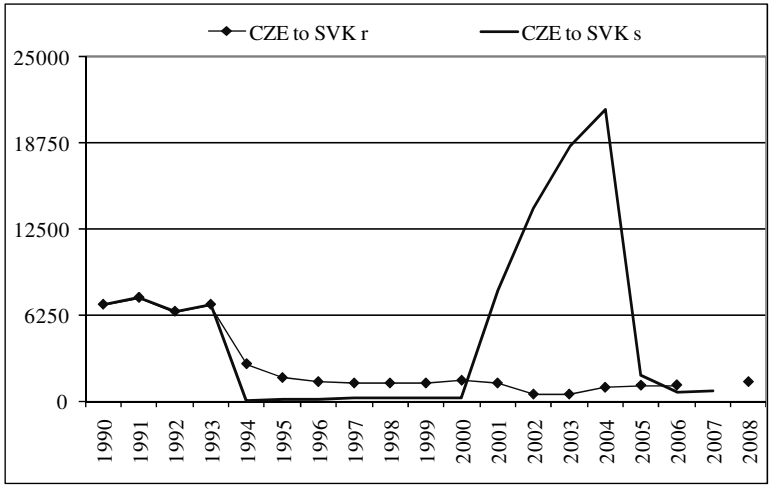


Figure 6: Migration between Slovak Republic and Czech Republic
r – data according to the receiving countries s – data according to the sending countries
Source: EUROSTAT, table migr_emi3nxt and table migr_imm5prv

A very low level of both immigration and emigration is reported by Slovakia and Poland during the whole period for which the data are available and it does not allow the identification of the changes in the flow magnitude observed by the partner country.
The flows among Czech Republic and Slovak Republic are the same until 1993, when former Czechoslovakia was split to two separated countries. After the dissolution of Czechoslovakia on 1 January 1993 the previously internal movements between the territories of the Czech Republic and Slovakia became international migration flows. In 2002 and 2003 flows from Slovakia and Poland are average 41 times higher than those reported by the sending countries. As concerns the sudden jumps observed in the Czech data, they might be explained by the changes in the definitions. In the Czech Republic until 2000 the statistics covered permanent migration only, as registered in the population register, similarly to Poland and Slovakia.

Since 2001, data from the aliens register were used as well: immigration statistics covered persons who stayed over one year (the exact criteria varied over time) and emigration statistics included data on permits that expired, in addition to self-reported departures for permanent stay abroad.

3. Conclusion

Achieving comparability of international migration statistics is a difficult task. The legislation and administrative procedures concerning registration, which is the main source of information on migration flows in the selected countries, will continue to differ. It should be noted that the lack of comparability of statistics on international migration flows is strictly linked with the lack of comparability of statistics on population stocks, so both problems should be solved simultaneously.

And at the end some recommendations for end users of the data of international migration:

- try to find out what is the real content of the data
- do not rely on one source
- do not draw conclusion without taking the definition into account.

4. Sources

[1]EUROSTAT: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

[2]ČESKÝ STATISTICKÝ ÚŘAD: www.czso.cz

[3]ŠTATISTICKÝ ÚRAD SLOVENSKEJ REPUBLIKY: www.statistics.sk

[4]Hungarian central statistical office: www.portal.ksh.hu

[5]Central statistical Office of Poland: www.stat.gov.pl

Adresa autora:
Eva Kačerová, RNDr.
Katedra demografie FIS VŠE v Praze
Nám. W. Churchilla 4, 130 67 Praha 3
kacerova@vse.cz

Regionálna zhluková analýza priemernej mesačnej mzdy na Slovensku
Regional Cluster Analysis of Average Monthly Salary in Slovakia

Samuel Korony

Abstract: Development and status of regions can be expressed by various indicators. One of the most important is the average salary. In the paper available regional district data from the years 2001 and 2008 (online database Regdat of Statistical Office of the Slovak Republic) are analyzed by regression and cluster analyses. Both, north-south and east-west regional differentiation is evident either from graphs or regression analysis results in case of the average salary. The plot of relative variance shows optimum number of clusters equal to three. What is more it has got nice geographic interpretation.

Key words: Exploratory data analysis, Data mining, Cluster analysis, Regional analysis.

Kľúčové slová: exploračná analýza dát, data mining, zhluková analýza, regionálna analýza.

JEL classification: C21, D31, R11.

1. Úvod

Jeden z najdôležitejších syntetických ukazovateľov vyspelosti regiónov je priemerná mzda (Korec 2005). Na jej analýzy je najjednoduchšie použiť údaje o priemernej mesačnej mzde, ktoré sú voľne prístupné v databáze Regdat na stránke Štatistického úradu SR.

V rámci sledovania vývoja hospodárstva SR sa zisťuje zamestnanosť a mzdy podľa podnikového výkazníctva. Údaje sú spracované z ročného výkazu, ktorý vypĺňajú všetky podniky a organizácie s 20 a viac zamestnancami (za organizácie peňažníctva a poisťovníctva bez ohľadu na počet zamestnancov). Zamestnaní sú vykázaní v tom okrese, kde skutočne pracujú (teda majú svoje skutočné pracovisko a nielen sídlo podniku). Medzi zamestnancov sa nezahŕňajú osoby na materskej (rodičovskej) dovolenke a ozbrojené zložky. Počty zamestnancov sa uvádzajú vo fyzických osobách. Priemerná mesačná mzda (podnikové výkazníctvo) zahŕňa čiastku mzdových nákladov, vyplácanú vlastným zamestnancom ako odmenu za prácu alebo jej náhradu na základe právneho vzťahu (pracovného, služobného, štátnozamestnaneckého alebo členského pomeru) k zamestnávateľovi. Je to hrubá mzda, nezníženú o zákonné alebo so zamestnancom dohodnuté zrážky (údaje sú prepočítané na fyzické osoby a sú bez podnikateľských príjmov).

Pre účely analýzy sme zobrali údaje o priemernej mesačnej mzde v Eurách v roku 2001 a 2008. Zmyslom výberu daných znakov bolo zistiť, ako sa zmenila priemerná mesačná mzda v roku 2008 v porovnaní s rokom 2001 v absolútnej aj relatívnej mierke. A tiež aká bola situácia v jednotlivých okresoch v oboch porovnávaných rokoch.

V ďalšej časti si stručne uvedieme dôvody pre výber analyzovaných jednotiek.

2. Voľba štatistických jednotiek

V našej aplikácii štatistickej analýzy na regionálne dáta sme použili priestorové štatistické jednotky na úrovni okresov (NUTS 4 alebo LAU 1), pretože údaje na úrovni krajov (NUTS 3) alebo ešte vyššie na úrovni oblastí (NUTS 2) už skryjú akékoľvek konkrétne rozdiely medzi regiónmi na Slovensku. Nanešťastie sa práve údaje a ich analýzy na úrovni NUTS 3 alebo NUTS 2 prezentujú v publikáciách OECD.

S počtom zvolených priestorových jednotiek súvisí aj problém štatistického testovania. Z pohľadu klasickej štatistiky sa nesmú robiť štatistické testy na základné súbory.

V moderných publikáciách sa používajú testy aj na základné súbory, ktoré nie sú príliš veľké (maximálny celkový počet niekoľko sto jednotiek). Z tohto pohľadu je okres ako priestorová jednotka SR v istom zmysle optimálny. Nie je ich príliš veľa (ako napr. zhruba 3000 obcí na Slovensku), kedy by sa označili za signifikantné aj zanedbateľné rozdiely. Na druhej strane okresy poskytujú určité informácie o stave vo vnútri krajov.

Pred výsledkami analýzy údajov uvedieme metódy a spôsob ich použitia pri analýze dát.

3. Použité metódy

Ako zhlukovú analýzu sme použili metódu *k*-priemerov s euklidovskou vzdialenosťou a hodnotami štandardizovanými smerodajnou odchýlkou a lineárnu diskriminačnú analýzu pre spätné overenie zhlukov. Pre zistenie geografickej distribúcie hodnôt (závislosti priemernej mesačnej mzdy od zemepisnej šírky a dĺžky) to bola lineárna regresná analýza. Numerické výpočty a zobrazenie výsledkov analýzy sme urobili v štatistických systémoch SPSS 18, NCSS 2001 a MYSTAT 12.

Pri výbere znakov sme zobrali do úvahy hlavný cieľ: či sa regionálne disparity zväčšujú alebo zmenšujú z pohľadu priemernej mesačnej mzdy.

S cieľom geografickej interpretácie polohy nájdených zhlukov na území SR sa nám po vyskúšaní niekoľkých geografických projekcií zo systému MYSTAT osvedčila Mercatorova konformná projekcia so zadanými zemepisnými súradnicami okresov (presnejšie ich okresných miest), ktorá zachováva správny pomer medzi vertikálne a horizontálne zobrazenými znakmi. Mestské bratislavské a košické okresy sme zobrazili pod seba, aby sa ich symboly neprekrývali.

Prvým krokom pri analýze dát by malo byť zistenie základných vlastností skúmaných znakov a objektov.

4. Exploračná analýza dát

Základné štatistické parametre skúmaných znakov sú v tabuľke 1. Z nej vyplýva, že všetky parametre sa zväčšili: aritmetický priemer miezd zo všetkých slovenských okresov z 385 Eur na 695 Eur ($p < 0{,}001$), smerodajná odchýlka zo 74 Eur na 154 Eur, variačný koeficient z 19 percent na 22 percent, minimum z 294 Eur na 502 Eur, maximum zo 657 Eur na 1 352 Eur. Platí to aj pre rozsah (z 363 Eur na 850 Eur).

Tabuľka 1: Štatistické parametre priemernej mesačnej mzdy v Eurách v roku 2001, 2008 a jej absolútneho a relatívneho prírastku

Znak	Priemer	Smer. odch.	Var. koef.	Minimum	Maximum	Rozsah
Mzda 2001	384.58	74.2	0.193	293.87	656.84	362.97
Mzda 2008	695.1	154.39	0.222	502.48	1352.2	849.72
Abs. prírastok	310.52	85.83	0.276	186.41	695.79	509.38
Rel. prírastok	0.802	0.111	0.139	0.588	1.183	0.595

V roku 2001 bola minimálna hodnota mzdy v okrese Sobrance, v roku 2008 to bolo v okrese Bardejov. Maximálna bola v okrese Bratislava I v roku 2001 a Bratislava II (rok 2008).

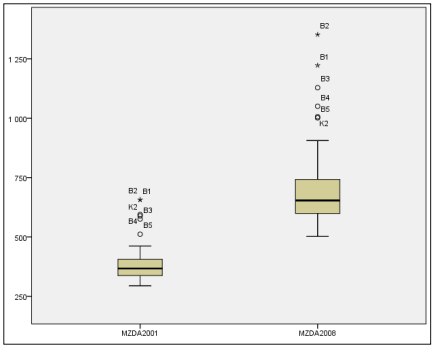
Absolútny prírastok ako rozdiel mzdy v roku 2008 a mzdy v roku 2001 mal priemernú hodnotu 311 Eur. Najmenší absolútny prírastok bol v okrese Bardejov - 186 Eur, najväčší v okrese Bratislava II - 696 Eur. Relatívny prírastok mal priemernú hodnotu 80 %. Ak uvažujeme priemernú mesačnú mzdu za celé Slovensko, potom relatívny prírastok v roku 2008 voči roku 2001 je 791,78 Eur / 423,69 Eur = 1,869 (87 %). Najmenší relatívny prírastok

bol v okrese Púchov a Bardejov 59 %. Najväčší relatívny prírastok mesačnej mzdy medzi rokmi 2001 a 2008 je v okrese Kysucké Nové Mesto (118 %). Viac ako 100 % dosahoval prírastok v ďalších dvoch okresoch: Sobrance 108 % a Bratislava II 106 %.

Ak zoberieme za základ priemernú mesačnú mzdu za celé Slovensko, tak výsledok porovnania je oveľa horší. V roku 2001 bolo 67 okresov s mzdou menšou ako bola priemerná mesačná mzda za celé Slovensko (423,69 Eura), v roku 2008 to bolo 68 okresov (791,78 Eur). To je podiel 85 resp. 86 percent! V roku 2001 mali nasledovné okresy priemernú mesačnú mzdu vyššiu ako bol celoslovenský priemer: Žiar nad Hronom, Malacky, Banská Bystrica, Trnava, Púchov, Košice I a II, a všetky bratislavské okresy I-V. Za rok 2008 to boli okresy Kysucké Nové Mesto, Žilina, Malacky, Trnava a opäť Košice I a II a bratislavské okresy. Jednoznačne sa potvrdzuje dominancia všetkých piatich mestských bratislavských okresov a dvoch košických. Na základe samotnej priemernej mesačnej mzdy nie je ťažké identifikovať podstatný vplyv veľkých zamestnávateľov (napr. automobilového priemyslu) a s ním súvisiacich odvetví (napr. hutníctva a strojárstva) na Slovensku.

Pre objektívne posúdenie, či sa rozptyl priemernej mesačnej mzdy v slovenských okresoch zmenil od roku 2001 do roku 2008, sme použili Levenov robustný test, ktorý testuje absolútne odchýlky hodnôt od mediánu. Na základe uvedeného testu môžeme skonštatovať, že sa rozptyl priemernej mesačnej mzdy v slovenských okresoch medzi rokmi 2001 a 2008 zväčšil významne (p = 0,0005).

Na grafe 1 sú zobrazené boxploty priemernej mesačnej mzdy v Eurách. Je na nich vidieť, že v roku 2001 odľahlé okresy v zmysle veľkých hodnôt mesačnej mzdy sú všetky bratislavské okresy Bratislava I až V a jeden košický okres Košice II. Situácia je podobná v roku 2008.

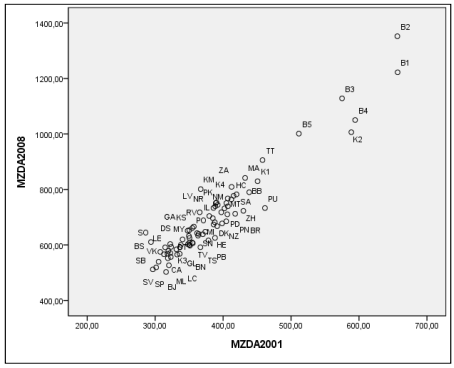


Graf 1: Boxploty priemernej mesačnej mzdy za okresy v rokoch 2001 a 2008 (Euro)

Na grafe 2 je bodový graf empirickej závislosti priemernej mesačnej mzdy v Eurách v roku 2008 od mzdy v roku 2001. Na prvý pohľad je zrejmé, že mzda sa zvýšila. Ak by sa nezmenila podstatne, tak by hodnoty boli sústredené okolo priamky $y = x$, ktorá má sklon 45 stupňov a intervaly hodnôt na osiach by boli rovnaké. Z grafu je navyše vidieť, že odľahlosť všetkých bratislavských mestských okresov a okresu Košice II zrejmá z boxplotov, ostáva aj na dvojrozmernom zobrazení.

Všetky bratislavské okresy Bratislava I-V a jeden košický okres Košice II v roku 2001 presahovali úroveň mzdy za niektoré okresy v roku 2008. Pre porovnanie zoberieme najvyššiu mzdu za rok 2001, ktorá bola v okrese Bratislava I na úrovni 656,84 Eura.

Potom v roku 2008 bolo spolu presne 40 (polovica) slovenských okresov so mzdou menšou ako bola mzda za rok 2001 v okrese Bratislava I! A to nezohľadňujeme infláciu, ktorá kumulatívne medziročne od roku 2001 do roku 2008 dosiahla hodnotu 39 percent!

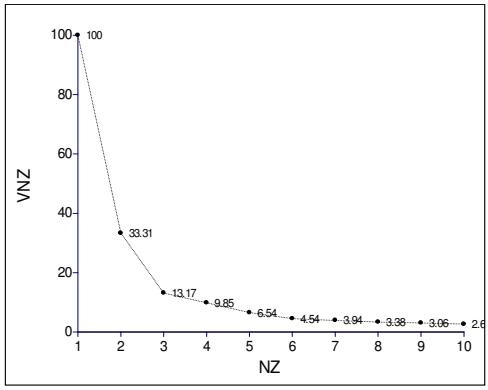


Graf 2: Empirická závislosť priemernej mesačnej mzdy v roku 2008 od mzdy v roku 2001

Po zistení základných charakteristík pristúpime k samotnej zhlukovej analýze priemernej mesačnej mzdy.

5. Zhluková regionálna analýza dát

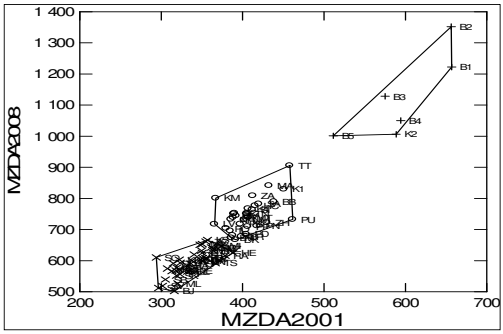
Na grafe 3 je zobrazený relatívny rozptyl pre zhlukovú analýzu (metóda *k*-priemerov) so znakmi priemernej mesačnej mzdy v Eurách v roku 2001 a v roku 2008. Z neho je vidieť, že relatívne najlepší počet zhlukov je tri.



Graf 3: Relatívny rozptyl pre zhlukovú analýzu priemernej mesačnej mzdy v r. 2001 a 2008

Na grafe 4 sú zobrazené tri zistené zhluky v priestore priemernej mesačnej mzdy roku 2008 a 2001. Zhluková metóda *k*-priemerov vydelila všetky bratislavské okresy I-V a okres Košice II do osobitného tretieho zhluku. Všetky zhluky sú dobre definované a navzájom oddelené.

V tabuľke 2 sú štatistické parametre troch identifikovaných zhlukov. Lineárna diskriminačná analýza má v tomto prípade nulovú klasifikačnú chybu.



Graf 4: Závislosť priemernej mesačnej mzdy roku 2008 od mzdy roku 2001 s tromi zhlukmi

Tabuľka 2: Štatistické parametre troch zhlukov priemernej mesačnej mzdy v r. 2001 a 2008

Znak	Parameter / zhluk	1	2	3
MZDA2001	**Priemer**	405.93	338.47	597.11
	Smer. Odchýlka	24.12	23.46	54.75
	Minimum	365.56	293.87	511.45
	Maximum	461.59	388.10	656.84
MZDA2008	**Priemer**	745.53	596.22	1126.77
	Smer. odchýlka	53.87	40.86	138.62
	Minimum	667.59	502.48	1001.26
	Maximum	905.81	666.20	1352.20
	N	31	42	6

V zhluku číslo dva je 42 okresov s najmenšou priemernou mesačnou mzdou 338 Eur za rok 2001 a 596 Eur v roku 2008. Ako je vidieť zo zemepisnej polohy na grafe 5 (symbol = x), ide o takmer všetky okresy východného Slovenska, južnej časti Banskobystrického, Trnavského a Nitrianskeho kraja a severnej časti Žilinského kraja.

V prvom zhluku (symbol = ○) je 31 ostatných okresov s vyššou priemernou mzdou 406 Eur (rok 2001) a 746 Eur (rok 2008).

Zostáva skupina šiestich okresov zo zhluku číslo tri (symbol = +), kde sú všetky mestské bratislavské okresy Bratislava I-V a jeden košický okres Košice II s priemerom 597 (2001) a 1127 Eur (2008).

Pre overenie, či priemerná mesačná mzda za jednotlivé okresy závisí priamo aj od ich geografickej polohy sme urobili lineárnu regresnú analýzu. Tá potvrdila závislosť pre mzdy za obidva skúmané roky. Všetky regresné parametre sú záporné a signifikantné.

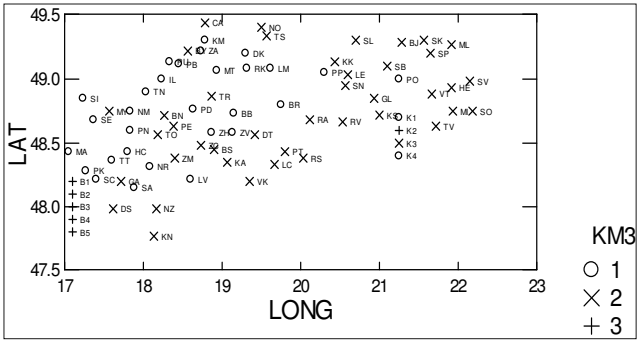
Pre závislosť priemernej mesačnej mzdy za rok 2001 sú regresné parametre -41,3 (p = 0,047) pre zemepisnú šírku a -16,7 (p = 0,004) pre zemepisnú dĺžku.

Ak je teda zemepisná šírka okresu väčšia o jeden stupeň, potom bola priemerná mesačná mzda v danom okrese menšia v roku 2001 v priemere o 41 Eur. Ak je zemepisná dĺžka okresu väčšia o jeden stupeň, potom bola priemerná mesačná mzda menšia v priemere o 17 Eur za rok 2001.

Pre rok 2008 sú parametre -92,3 (p = 0,028) pre zemepisnú šírku a -38,8 (p = 0,001) pre zemepisnú dĺžku.

To znamená, že pri zväčšení severnej zemepisnej šírky o jeden stupeň bola priemerná mesačná mzda menšia o 92 Eur v roku 2008!

Pre východnú zemepisnú dĺžku je to v priemere 39 Eur na každý ďalší stupeň. Bratislava leží v juhozápadnom cípe Slovenska a preto sa môžu dané zistené závislosti vztiahnuť na ňu ako na referenčný bod – počiatok.



Graf 5: Geografická poloha troch zhlukov okresov na základe priemernej mesačnej mzdy v roku 2001 a 2008

6. Záver

Na základe údajov o priemernej mesačnej mzde za jednotlivé okresy Slovenska je nutné konštatovať, že regionálne disparity v príjmoch obyvateľstva a teda aj životnej úrovni sa medzi rokmi 2001 a 2008 nezmenšili (čo je jedna z požiadaviek alebo odporúčaní EÚ a Svetovej banky), ale naopak ešte viac prehĺbili. Rok 2001 bol rok, keď sme ešte neboli v EÚ, ale už sme mali de jure niektoré disparity zmenšovať. Rok 2008 by už mal niesť známky vplyvu vstupu do EÚ (pozitívny ekonomický dopad) a tiež u nás začínajúcej ekonomickej krízy (negatívny dopad).

Na základe údajov sme zistili aj výraznú anomáliu vo forme signifikantnej závislosti priemernej mesačnej mzdy od polohy okresu. Priemerná mesačná mzda klesá smerom na sever a na východ. Bratislava má prakticky minimálnu zemepisnú dĺžku aj šírku. Ak sa posúvame smerom od Bratislavy ďalej na Slovensko, tak klesá priemerná mesačná mzda v obidvoch možných zemepisných smeroch na sever aj na východ.

7. Literatúra

KOREC, P. et al. 2005. Regionálny rozvoj Slovenska v rokoch 1989-2004 : Identifikácia menej rozvinutých regiónov Slovenska. 1. vyd. Bratislava : Geo-grafika, 2005. ISBN 80-969338-0-9

Adresa autora:

RNDr. Samuel Koróny, PhD., Email: samuel.korony@umb.sk
Ústav vedy a výskumu UMB, Cesta na amfiteáter 1, 974 01 Banská Bystrica

Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-0649-07.

Porovnanie krajín strednej a východnej Európy na základe hodnôt troch ukazovateľov stability banky

Comparison of CEE Countries Based on Values of Three Indicators of Bank Stability

Mária Kóšová

Abstract: Each bank in the world should work to make a profit. However, it should also allocate financial resources to optimize the risk, profit and liquidity. This article presents a comparison of 10 countries from Central and Eastern Europe in view of the three indicators of bank stability in the period 2006-2008. As a method of comparison is used cluster analysis.

Key words: Cluster analysis, Principal component analysis, Indicator of bank stability, Ward's method

Kľúčové slová: Zhluková analýza, metóda hlavných komponentov, ukazovateľ stability banky, Wardova metóda

JEL classification: C38

1. Úvod

Ekonomický rozvoj krajiny úzko súvisí s pohybom peňazí. Tento pohyb sprostredkovávajú banky. Aby banka pracovala čo najefektívnejšie, musí optimalizovať tri faktory – riziko, zisk a likviditu. Možno povedať, že likvidita je schopnosť banky vyplatiť svoje záväzky bez neprijateľných strát, teda mať dostatok hotovosti alebo aktív, ktoré sa dajú v prípade potreby použiť. Avšak zároveň platí, čím je vyššia likvidita banky, tým menší je jej zisk. Okrem iného aj zlá optimalizácia týchto faktorov v bankách USA viedla ku vzniku hospodárskej krízy vo svete v roku 2008. V našom príspevku budeme porovnávať 10 krajín strednej a východnej Európy z hľadiska troch ukazovateľov stability banky v období rokov 2006-2008. Teda v období troch rokov pred vypuknutím svetovej hospodárskej krízy.

2. Údaje a metodika

Zamerali sme sa na porovnanie krajín V4 (Česká republika, Poľsko, Slovenská republika, Maďarsko), Bulharska, Rumunska, Lotyšska, Litvy, Estónska a Slovinska. Údaje o bankách sme získali z databázy Bankscope a v prípade ukazovateľa podielu likvidných aktív k likvidným pasívam z výročných správ jednotlivých bánk. V rámci Českej republiky sme pracovali s údajmi z 15-tich bánk, v rámci Poľska z 19-tich bánk, Slovenskej republiky zo 14-tich, Maďarska z 15-tich, Bulharska z 19-tich, Rumunska z 20-tich, Lotyšska z 19-tich, Litvy z 8-mich, Estónska zo 7-mich, Slovinska zo 16-tich bánk. Sledované sú nasledovné tri ukazovatele ich stability (v percentách):

1. podiel hotovosti a obchodovateľných cenných papierov k celkovým aktívam (Cash and Due from Banks/ Total Assets)
2. podiel likvidných aktív so splatnosťou do 3 mesiacov k likvidným pasívam so splatnosťou do 3 mesiacov (Liquid Assets / Liquid Liabilities)
3. podiel vlastného kapitálu k celkovým aktívam (Equity/ Total Assets)

Zo získaných hodnôt ukazovateľov pre jednotlivé banky v rámci danej krajiny a roku sme ako celkovú hodnotu ukazovateľa pre danú krajinu za rok použili v prípade ukazovateľa Cash and Due from Banks/ Total Assets a ukazovateľa Liquid Assets / Liquid Liabilities ich priemer, v prípade tretieho ukazovateľa Equity/ Total Assets sme použili ich vážený priemer. Následne

sme aj tieto priemery pre konkrétne roky spriemerovali, čím sme získali celkovú hodnotu ukazovateľa (premennej) pre krajinu (pozri tabuľka 1).

Tabuľka 1: Vstupné dáta

	Cash & Due from banks/Total Assets (v %)	Liquid Assets / Liquid Liabilities (v %)	Equity / Total Assets (v %)
CZ	3,94	66,19	7,67
SK	6,47	59,55	7,53
HU	2,75	82	5,94
PL	4,12	78,67	9,74
BG	11,39	98	11,2
RO	21,16	136,33	9,5
LV	6,09	67	7,97
EE	7,32	100,67	9,43
LT	8,54	53,33	7,53
SLO	2,52	74	7,93

V tabuľke 2 je možné vidieť niektoré popisné štatistiky vstupných dát.

Tabuľka 2: Priemer a smerodajná odchýlka vstupných dát vzhľadom na jednu krajinu

	Mean	Std. Dev.
Cash & Due from banks/ Total Assets (v %)	7,43000	5,55080
Liquid Assets/ Liquid Liabilities (v %)	81,57400	24,56402
Equity/ Total Assets (v %)	8,44400	1,50522

Pre porovnanie krajín sme zvolili zhlukovú analýzu. Použitie zhlukovej analýzy si vyžaduje nezávislé premenné (ukazovatele). Závislosť resp. nezávislosť premenných sme overili vypočítaním koeficientu viacnásobnej korelácie. Výslednú korelačnú maticu možno vidieť v tabuľke 3.

Tabuľka 3: Korelačná matica

	Cash & Due from banks/Total Assets (v %)	Liquid Assets / Liquid Liabilities (v %)	Equity / Total Assets (v %)
Cash & Due from banks/ Total Assets (v %)	1,000000	0,751795	0,511408
Liquid Assets/ Liquid Liabilities (v %)	0,751795	1,000000	0,558178
Equity / Total Assets (v %)	0,511408	0,558178	1,000000

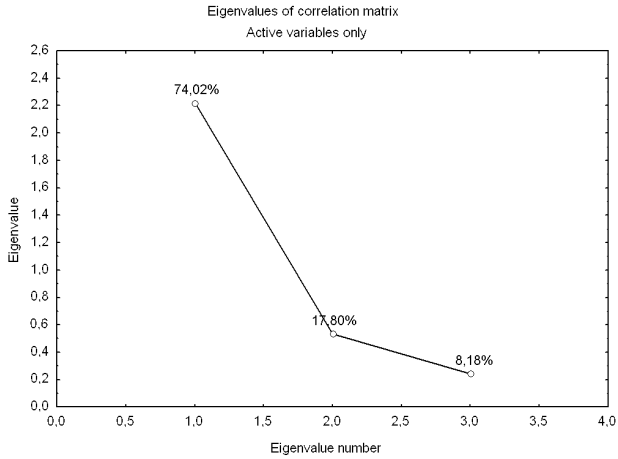
Keďže pôvodné premenné boli závislé, použili sme najskôr metódu hlavných komponentov (Principial Components & Classification Analysis v programe Štatistika). Prostredníctvom nej sme získali tri nezávislé hlavné komponenty (v programe Štatistika pomenované ako faktory). Tieto komponenty sú lineárnou kombináciou pôvodných premenných. Ako vidieť z tabuľky 4, prvý hlavný komponent Factor 1 je tvorený najväčšou váhou premennej Liquid Assets/ Liquid Liabilities, avšak aj ostatné dve premenné majú veľkú váhu. V druhom hlavnom komponente Factor 2 má najväčšiu váhu premenná Equity/ Total Assets. Táto má záporné znamienko, kým ostané dve premenné Cash and Due from Banks/ Total Assets a Liquid Assets/ Liquid Liabilities majú kladné znamienko. Tretí hlavný komponent Factor 3 je

tvorený najväčšou váhou premenných Cash and Due from Banks/ Total Assets a Liquid Assets/ Liquid Liabilities.

Tabuľka 4: Váhy premenných v hlavných komponentoch Factor 1, Factor 2, Factor 3

	Factor 1	Factor 2	Factor 3
Cash and Due from Banks/ Total Assets	-0,886255	0,321881	0,333085
Liquid Assets/ Liquid Liabilities	-0,904871	0,220803	-0,363943
Equity/ Total Assets	-0,785103	-0,617838	0,043464

Z obrázku 1 vyplýva, že prvý hlavný komponent vysvetľuje až 74,02% celkovej variability dát, druhý 17,80% celkovej variability dát a tretí len 8,18% celkovej variability dát.



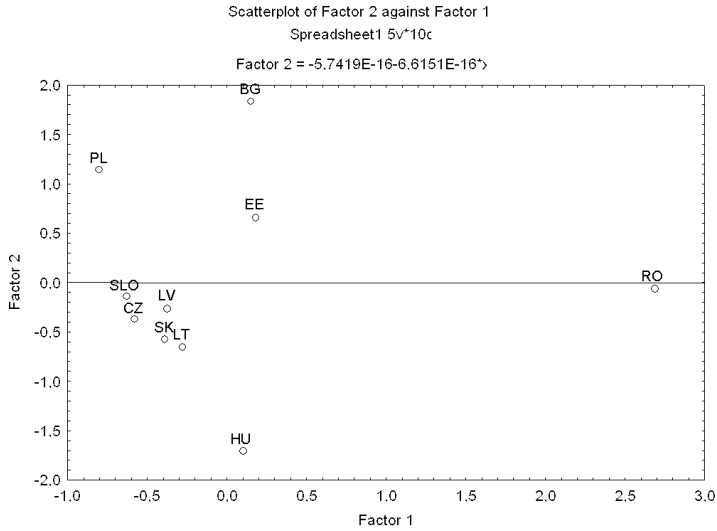
Obrázok 1: Sutinový graf vlastných čísel

Prvý a druhý komponent spolu vysvetľujú až 91,82% celkovej variability pôvodných dát. Vzhľadom na vysoké korelácie jednotlivých premenných s Factorom 1 sme použitím faktorovej analýzy v programe Štatistika extrahovali dva faktory, ktoré po použití rotácie Varimax raw (Varimax prostý) vytvárajú jednoduchú štruktúru uvedenú v tabuľke 5. Vidíme, že ukazovateľe Cash and Due from Banks/ Total Assets a Liquidy Assets/ Liquidity Liabilities vysoko korelujú s prvým extrahovaným faktorom a naopak s druhým faktorom korelujú nevýznamne. Tretí ukazovateľ Equity/ Total Assets vysoko koreluje s druhým extrahovaným faktorom, kým s prvým extrahovaným faktorom má nízku koreláciu. Prvý faktor možno nazvať faktor likviditného rizika.

Tabuľka 5: Váhy premenných vo faktoroch Factor 1 a Factor 2 s rotáciou Varimax raw

	Factor 1	Factor 2
Cash & Due from banks/ Total Assets (v %)	0,906911	0,258006
Liquid Assets/ Liquid Liabilities (v %)	0,862817	0,350845
Equity / Total Assets (v %)	0,274670	0,960556

Jednotlivé krajiny môžme zobraziť do grafu vzhľadom na tieto extrahované faktory s použitím rotácie Varimax raw (pozri obrázok 2).



Obrázok 2: Bodový graf krajín pre Factor1 a Factor2 s rotáciou Varimax raw

Ako extrémna hodnota (outlier) sa prejavilo Rumunsko, ktoré má najvyšší podiel hotovosti a obchodovateľných cenných papierov k celkovým aktívam (Cash and Due from Banks/ Total Assets) a najvyšší podiel likvidných aktív k likvidným pasívam (Liquid Assets/ Liquid Liabilities). Ďalšou extrémnou hodnotou je tiež Estónsko, ktoré malo druhý najvyšší podiel likvidných aktív k likvidným pasívam (Liquid Assets/ Liquid Liabilieties) a Poľsko, ktoré malo druhý najvyšší podiel vlastného kapitálu k celkovým aktívam (Equity/ Total Assets). Rovnako tak možno považovať za extrémnu hodnotu aj Bulharsko, ktoré malo najvyšší podiel vlastného kapitálu k celkovým aktívam (Equity/ Total Assets).

3. Zhluková analýza

Zhluková analýza patrí medzi metódy, ktoré sa zaoberajú vyšetrovaním podobnosti viacrozmerných objektov (t.j. objektov, u ktorých je zmerané väčšie množstvo premenných) a ich klasifikácii do tried teda zhlukov ([3]).

V analýze sme aplikovali meranie vzdialenosti medzi objektmi pomocou euklidovskej vzdialenosti, ktorá vyjadruje ich vzdušnú vzdialenosť. Výsledný dendrogram sme získali Wardovou metódou. Táto minimalizuje vnútrozhlukový rozptyl, teda celkový súčet štvorcov odchýlok všetkých hodnôt od príslušných zhlukových priemerov.

Ako vstupné dáta zhlukovej analýzy boli použité hodnoty všetkých troch hlavných komponentov Factor 1, Factor 2 a Factor 3 (pozri tabuľka 6).

Z dendrogramu na obrázku 4 je vidieť, že rezom na hladine vzdialenosti 1,5 dosiahneme rozdelenie krajín na päť zhlukov.

Zhluk č. 1: Litva, Slovenská republika, Lotyšsko, Česká republika a Slovinsko.

Tento zhluk tvoria krajiny, ktoré nadobúdali podpriemerné hodnoty všetkých troch ukazovateľov.

Zhluk č. 2: Maďarsko.

Ukazovateľ Liquid Assets/ Liquid Liabilities nadobúdal v Maďarsku priemernú hodnotu ostatné dva ukazovatele nadobúdali podpriemerné hodnoty.

Zhluk č. 3: Estónsko a Poľsko.

Estónsko a Poľsko sú krajiny, ktorých hodnoty prvého ukazovateľa boli podpriemerné a tretieho ukazovateľa mierne nadpriemerné.

Zhluk č. 4: Bulharsko.

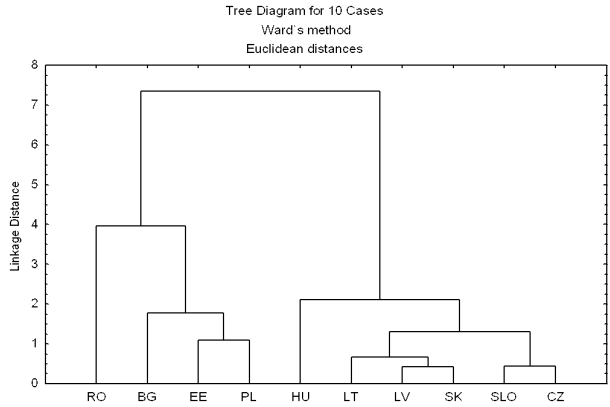
Bulharsko nadobudlo nadpriemerné hodnoty všetkých troch ukazovateľov. Tiež má zároveň najväčšiu hodnotu ukazovateľa Equity/ Total Assets.

Zhluk č. 5: Rumunsko.

Rumunsko nadobúdalo vysoko nadpriemerné hodnoty prvého a druhého ukazovateľa.

Tabuľka 6: Hodnoty hlavných komponentov Factor 1, Factor 2 a Factor 3

	Factor 1	Factor 2	Factor 3
CZ	1,02514	-0,03142	-0,007757
SK	0,96721	0,16629	0,489254
HU	1,36734	1,04027	-0,725764
PL	-0,02719	-1,02626	-0,238605
BG	-1,79498	-1,03166	0,149086
RO	-3,19427	1,16983	0,087049
LV	0,66975	-0,01936	0,245997
EE	-0,80538	-0,32764	-0,527106
LT	0,89918	0,25403	0,926130
SLO	0,89321	-0,19407	-0,398285



Obrázok 4: Dendrogram pre 10 krajín

4. Záver

Zaujímavým výsledkom je spojenie Slovenskej republiky s Lotyšskom, to znamená, že vzdialenosť medzi nimi je veľmi malá a teda sú bankové ukazovatele Slovenskej republiky najviac podobné s bankovými ukazovateľmi v Lotyšsku. Práve Lotyšsko bolo však po vypuknutí hospodárskej krízy jednou z krajín, v ktorých mala táto kríza veľmi negatívne následky na ekonomickú situáciu. Celkovo najpodobnejšími krajinami vzhľadom na tieto ukazovatele sú Česká republika a Slovinsko, ktoré nadobúdali výrazne podpriemerné hodnoty ukazovateľa Cash and Due from Bank/ Total Assets. Tiež už spomínané spojenie Slovenskej republiky a Lotyšska. Nasleduje pripojenie Litvy ku Slovenskej republike a Lotyšsku. Tieto tri krajiny mali hodnoty prvého s tretieho ukazovateľa blízke priemernej hodnote. Avšak k pripojeniu Českej republiky ku Slovenskej republike dochádza až vo štvrtom cykle. Ďalšie spojenia je možné vidieť na obrázku 3. Zaujímavé tiež je, že ak sa zameriame na rozmiestnenie krajín V4, zistíme, že tieto nielenže nemajú medzi sebou malé vzdialenosti, ale nie sú ani v spoločnom zhluku.

5. Literatúra

[1] HEBÁK, P. A KOL. 2007. Vícerzměrné statistické metody (3). Praha: Informatorium, spol. s r. o., 2007. 271s. ISBN 978-80-7333-001-9

[2] MEJSTŘÍK, M – PEČENÁ, M. – TEPLÝ, P. Bankovnictví - Basic principles of banking. 1. vyd. Praha : Karolinum , 2008. 627 s. ISBN 978-80-246-1500-4.

[3] MELOUN, M. – MILITKÝ, J. 2002. Kompendium statistického zpracování dat. Praha: Academia, 2002. 764 s. ISBN 80-200-1008-4.

[4] STANKOVIČOVÁ, I. – VOJTKOVÁ, M. 2007. Viacrozmerné štatistické metódy s aplikáciami. Bratislava: Iura Edition, 2007. 261s. ISBN 978-80-8078-152-1.

[5] ECB: Financial Stability Review, June 2010. Dostupné na internete:

http://www.ecb.int/pub/pdf/other/financialstabilityreview201006en.pdf?53cbe8ea87c0d0b4b154e37e1e461133

Adresa autora:

Mária Kóšová, Mgr.
Petzvalová 15
949 01 Nitra
maria.kosova@ukf.sk

Statistické vlastnosti bodového odhadu společného efektu léčby
Statistical Properties of an Overall Treatment Effect Point Estimator

Pavla Krajíčková

Abstract: One of aims of the meta-analysis of clinical trials is to determine the effectivity of a new type of treatment. For binary data, the efficiency can be measured by a probability of successful treatment. The paper presents statistical properties of an overall treatment effect point estimator in the meta-analysis based on multicentre trials. The construction of this estimator is suggested by Wimmer & Witkovsky (2004).

Abstrakt: Jedním z cílů meta-analýzy klinických pokusů je určení účinnosti nového typu léčby. V případě binárních dat může být tato účinnost měřena pomocí pravděpodobnosti úspěšné léčby. Příspěvek pojednává o některých statistických vlastnostech bodového odhadu společného efektu léčby v meta-analýze klinické studie provedené ve vícero zdravotnických zařízeních. Odvození tohoto bodového odhadu bylo navrženo Wimmer & Witkovsky (2004).

Key words: Multicenter Trial, Point Estimator, Probability of Success, Linear Model with Random Effects

Klíčová slova: Multicentrální studie, Bodový odhad, Pravděpodobnost úspěchu, Lineární model s náhodnými efekty

JEL classification: C13.

1. Introduction

Let us consider a clinical trial performed in I centers. Suppose that the number of subjects included in the trial in the ith center is $n_{T,i}$ for $i = 1, 2, \dots, I$. Patients in the ith center succeed with probability $p_{T,i}$ for $i = 1, 2, \dots, I$. All subjects are considered to be independent. One of the principal questions is: How is the efficiency of a new type of treatment?

Number of successes in the ith center is denoted by random variable $X_{T,i}$. Then $X_{T,i}$ has binomial distribution with the sample of size $n_{T,i}$ and the probability of success $p_{T,i}$. Random variables $X_{T,1}, \dots, X_{T,I}$ are stochastic independent. As Wimmer & Witkovsky (2004) suggested we will next work with random variables

$$\mathrm{Y_{T,i}} = \frac{X_{T,i}}{n_{T,i}} \tag{1}$$

for $i = 1, 2, \dots, I$.

2. The model and the overall treatment effect point estimator

Now suppose that the true probabilities of success in the ith center $p_{T,i}$ randomly fluctuate around common probabilities of success p_T and want to estimate the common probability of successful treatment p_T. So

$$p_{T,i} = p_T + b_{T,i} \tag{2}$$

for $i = 1, 2, \dots, I$ where $b_{T,i}$ is a random effect of the ith center and suppose that $b_{T,i} \sim N(0, \sigma^2_{T,0})$ which means $b_{T,i}$ is normally distributed with the mean 0 and the variance $\sigma^2_{T,0}$. [1]

As Wimmer & Witkovsky (2004) showed the final situation then can be represented by the linear model with random effects

[1] Of course it is supposed that $\sigma^2_{T,0}$ is such that "practically" $0 < p_T + b_{T,i} < 1$. In simulations it is ensured with a proper choice of $\sigma^2_{T,0}$.

$$Y_{T,i} = p_T + b_{T,i} + \varepsilon_{T,i} \quad (3)$$

for $i = 1, 2, \dots, I$ where $\varepsilon_{T,i}$ are error terms and $\varepsilon_{T,i} \sim N(0,\ \sigma^2_{T,i}/n_{T,i})$.

Wimmer & Witkovsky (2004) also suggested a point estimator of an overall treatment effect

$$\hat{p}_T = \frac{\sum_{i=1}^{I} \frac{x_{T,i}}{n_{T,i}\hat{\sigma}^2_{T,0} + \hat{\sigma}^2_{T,i}}}{\sum_{i=1}^{I} \frac{n_{T,i}}{n_{T,i}\hat{\sigma}^2_{T,0} + \hat{\sigma}^2_{T,i}}}. \quad (4)$$

where the unknown variance components $\sigma^2_{T,0}$ and $\sigma^2_{T,i}$ were replaced by their estimators $\hat{\sigma}^2_{T,0}$ and $\hat{\sigma}^2_{T,i}$, $i = 1, 2, \dots, I$. For estimation of $\sigma^2_{T,i}$ they used estimators proposed by Agresti & Caffo (2000):

$$\hat{\sigma}^2_{T,i} = \frac{X_{T,i}+2}{n_{T,i}+4}\left(1 - \frac{X_{T,i}+2}{n_{T,i}+4}\right) \quad (5)$$

for $i = 1, 2, \dots, I$.

As an estimator of $\sigma^2_{T,0}$ the Mandel & Paule (1982) procedure has been used. The estimator $\hat{\sigma}^2_{T,0}$ we obtain as an iterative solution of the following equations

$$\hat{\mu}_T^{MP} = \frac{\sum_{i=1}^{I} \frac{X_{T,i}}{n_{T,i}\hat{\sigma}^2_{T,0} + \hat{\sigma}^2_{T,i}}}{\sum_{i=1}^{I} \frac{n_{T,i}}{n_{T,i}\hat{\sigma}^2_{T,0} + \hat{\sigma}^2_{T,i}}}$$

$$\sum_{i=1}^{I} \frac{\left(\frac{X_{T,i}}{n_{T,i}} - \hat{\mu}_T^{MP}\right)^2}{\hat{\sigma}^2_{T,0} + \frac{\hat{\sigma}^2_{T,i}}{n_{T,i}}} = I - 1. \quad (6)$$

3. Simulation results

To explore the behavior of the overall treatment effect point estimator the simulations were conducted for two main different situations. In both situations the values of unknown

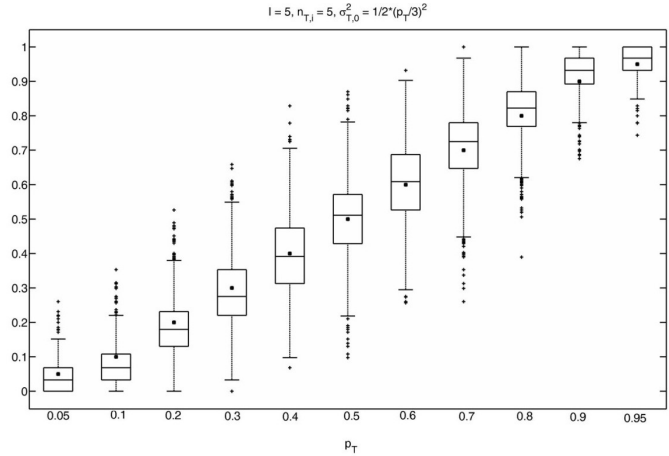


Figure 1: Balanced situation - $I = 5$ and $n_{T,i} = 5$

parameters I, $n_{T,i}$ and p_T were following: the numbers of center $I \in \{5, 10, 15, 20\}$, the numbers of subjects $n_{T,i} \in \{100,\ 50,\ 30,\ 15,\ 10,\ 5\}$, the variance of random effects was

for $p_T \leq 0.5$ $\sigma^2_{T,0} \in \left\{0,\ \frac{1}{4}\left(\frac{p_T}{3}\right)^2,\ \frac{1}{2}\left(\frac{p_T}{3}\right)^2, \frac{3}{4}\left(\frac{p_T}{3}\right)^2,\ \left(\frac{p_T}{3}\right)^2\right\}$ and for $p_T > 0.5$ $\sigma^2_{T,0} \in \left\{0,\ \frac{1}{4}\left(\frac{1-p_T}{3}\right)^2,\ \frac{1}{2}\left(\frac{1-p_T}{3}\right)^2, \frac{3}{4}\left(\frac{1-p_T}{3}\right)^2,\ \left(\frac{1-p_T}{3}\right)^2\right\}$ respective, common probability of success $p_T \in \{0.05,\ 0.1, 0.2, \ldots, 0.8, 0.9, 0.95\}$. For each situation 5000 replications were made.

In the first simulated situation (balanced situation) the number of subjects in the ith center $n_{T,i}$ was the same as the number of subjects in the jth center $n_{T,j}$ for all $i, j = 1, 2, \ldots, I$. In the second situations (unbalanced situation) the number of subjects in the ith center $n_{T,i}$ was not the same as the number of subjects in the jth center $n_{T,j}$ for $\neq j$, $i, j = 1, 2, \ldots, I$.

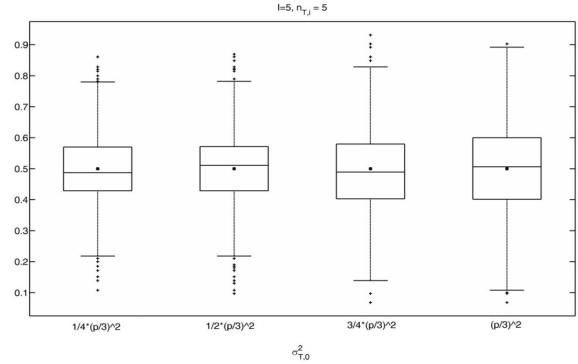


Figure 2: Balanced situation –different $\sigma^2_{T,0}$

Figure 1 show box plots of $\hat{p}_T$ in balanced situations where the simulation settings of parameters is following: $I = 5$, $n_{T,i} = 5$ and $\sigma^2_{T,0} = 1/2\,(p_T/3)^2$. Black squares at figures label the true value of p_T. As one can observe, for $p_T < 0.5$ the point estimator underestimate the true value of p_T and for $p_T > 0.5$ the point estimator overestimate the true value.

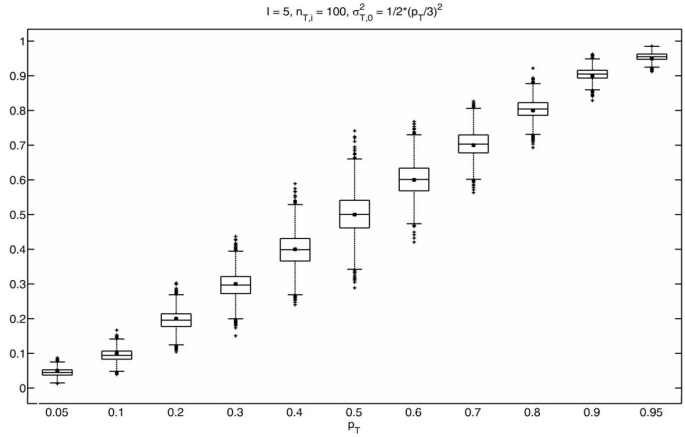


Figure 3: Balanced situation - $I = 5$ ***and*** $n_{T,i} = 100$

As p_T tends to 0.5, the variability of $\hat{p}_T$ grows. This is connected with growing value of $\sigma^2_{T,0}$ which is function of p_T. For fixed value of p_T and growing value of $\sigma^2_{T,0}$ the variability

of $\hat{p}_T$ grows too as is shown at figure 2. For all simulated balanced situations the difference in variability of $\hat{p}_T$ for different values of $\sigma^2_{T,0}$ was not greater than 0.02.

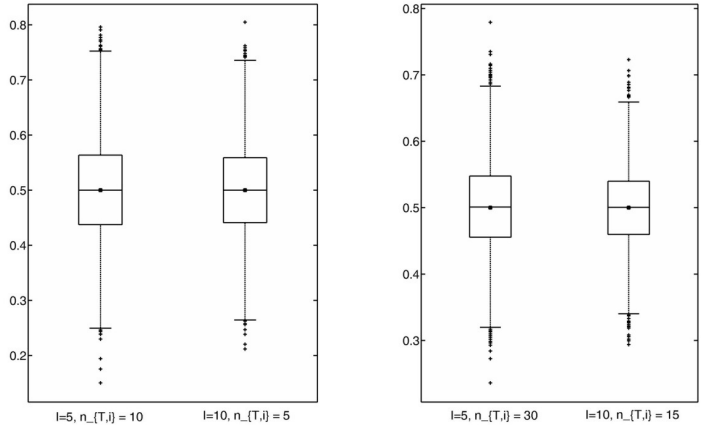


Figure 4: Balanced situation - comparison

As can be expected the variability of $\hat{p}_T$ decline with growing number of subjects in centers. This is illustrated at figure 3.

Now consider different balanced situations with the same total number of patients. Figure 4 show comparison for $I = 5$, $n_{T,i} = 10$ and $I = 10$, $n_{T,i} = 5$ and also for $I = 5$, $n_{T,i} = 30$ and $I = 10$, $n_{T,i} = 15$. Although the total number of subjects is same the variability of $\hat{p}_T$ is greater for smaller number of centers I.

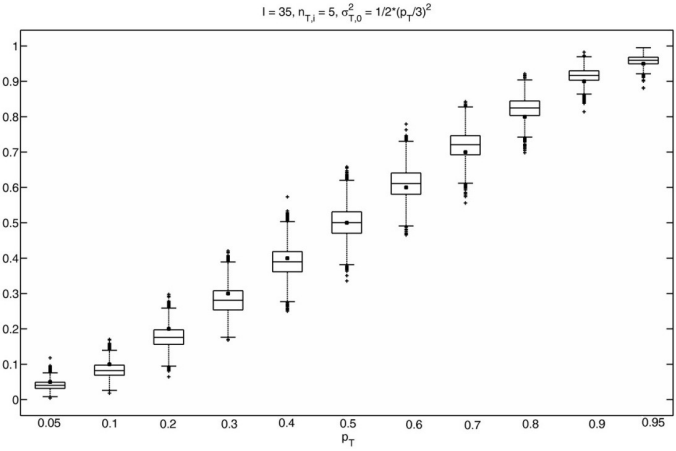


Figure 5: Balanced situation – $I = 35$ and $n_{T,i} = 5$

Figure 5 illustrate the situation for larger number of centers I, here for $I = 35$, $n_{T,i} = 5$. Due to greater total amount of subjects in the trial, the variability of $\hat{p}_T$ is smaller than in situations showed at figure 1. For smaller and higher values of p_T, for example $p_T = 0.1$ and $p_T = 0.9$, the true value of p_T come close to upper and lower edges of the box (figure 5).

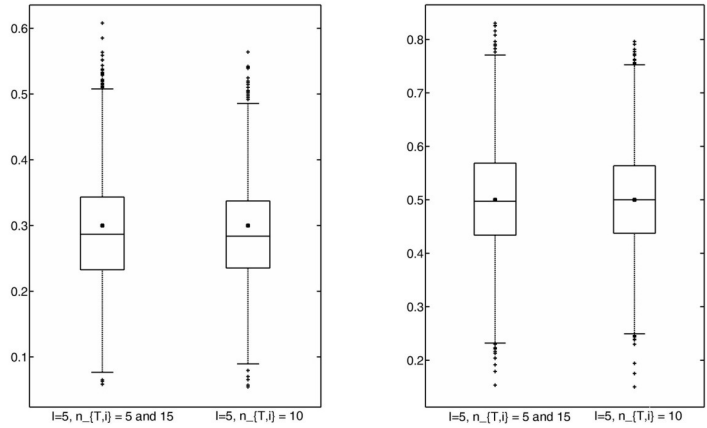


Figure 6: Balanced and unbalanced situations –comparison for $\boldsymbol{p_T \in \{0.3, 0.5\}}$

Now look at both the balanced and unbalanced situations. At figure 6 there are comparisons of $\hat{p}_T$ box plots for the same total number of patients but the first situation is unbalanced where $n_T = [5\ 15\ 5\ 15\ 5]'$ and the second is balanced where $n_T = [10\ 10\ 10\ 10\ 10]'$. As one can expected unbalanced situations have greater variability of $\hat{p}_T$.

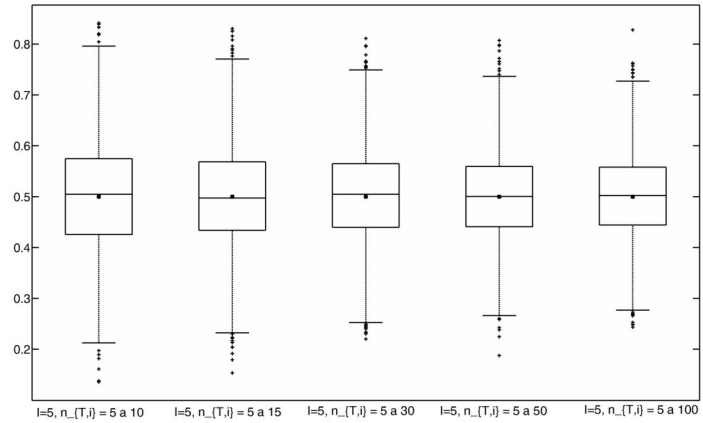


Figure 7: Unbalanced situations –comparison

The last figure 7 is also focused on changes in variability in unbalanced situations. As the number of patients in center grows the variability decline regardless of growing difference of subjects' number between centers.

For unbalanced situations the difference in variability of $\hat{p}_T$ for different values of $\sigma^2_{T,0}$ grows according increasing difference in numbers of subjects including in the trial in each center. For example the difference in variability of $\hat{p}_T$ for different values of $\sigma^2_{T,0}$ was 0.026 for $n_T = [5\ 10\ 5\ 10\ 5]'$, 0.041 for $n_T = [5\ 30\ 5\ 30\ 5]'$ and 0.054 for $n_T = [5\ 100\ 5\ 100\ 5]'$.

4. Conclusion

This paper was focused on some statistical properties of an overall treatment effect point estimator in the meta-analysis based on multicentre trials. The conducted simulation study suggests good practical properties of point estimator (4) proposed by Wimmer & Witkovsky (2004) but one have to be careful at boundary common probabilities of successful treatment.

5. Reference

[1]AGRESTI, A. – CAFFO, B. 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. In: The American Statistician, vol. 54, 2000.

[2]MANDEL, J. – PAULE R.C. 1982. Consensus values and weighting factors. In: Journal of Research of the National Bureau of Standards, vol. 87, 1982.

[3]WIMMER, G. – WITKOVSKÝ V. 2004. Konfidenčne intervaly pre efekt ošetrenia v klinických pokusoch. In: ROBUST 2004. Sborník prací 13. letní školy JČMF ROBUST 2004 uspořádané Jednotou českých matematiků a fyziků za podpory KPMS 2004 v Třešti. 2004, Praha: JČMF, 2004.

Author's address:

Pavla Krajíčková, Mgr.
Ústav matematiky a statistiky
Přírodovědecká fakulta Masarykovy Univerzity
Kotlářská 2, 611 37 Brno
pavla@krajickova.cz

Miery príjmovej nerovnosti
Income inequality measures

Viera Labudová

Abstract: Inequality is a broader concept than poverty in that it is defined over the entire population, and does not only focus on the poor. Among the most common metrics used to measure inequality are the Lorenz curve and the Gini index (Gini coefficient). Income inequality measures such as the generalised entropy index and the Atkinson index offer the ability to examine the effects of inequalities in different areas of the income spectrum.

Key words: Income inequality, generalized entropy, Theil's L index, Theil's T index, Atkinson index.

Kľúčové slová: Príjmová nerovnosť, generalizovaná entropia, Theilov L index, Theilov L index, Atkinsonov index .

JEL classification: D63, I32

1. Úvod

Nerovnosť príjmov a z neho vyplývajúce štatistické rozloženie domácností podľa príjmov je indikátorom ekonomickej vyspelosti krajiny, funkčnosti sociálnej politiky štátu a správnosti opatrení aplikovaných v sociálnej oblasti. Kvantifikácia nerovnosti príjmov býva často súčasťou širšej analýzy týkajúce sa chudoby a blahobytu, hoci existuje výrazný rozdiel medzi týmito pojmami.

Nerovnosť je širší pojem ako chudoba, pretože je definovaná a meraná na celej distribúcii príjmov (pri jej kvantifikácii sú vstupmi príjmy celej populácie), pričom chudoba je definovaná len na distribúcii pod určitou hranicou chudoby. Na druhej strane je nerovnosť oveľa užší pojem ako blahobyt. Hoci sú obidva koncepty založené na celej distribúcii príjmov, miery nerovnosti sú nezávislé od priemeru príjmového rozdelenia.

Na meranie príjmovej nerovnosti a pozorovanie jej zmien možno použiť rôzne miery, z nich najjednoduchšími sú miery variability, ako je rozptyl, smerodajná odchýlka, variačný koeficient. Ich nevýhodou pri meraní nerovnosti je to, že nie sú zhora ohraničené (Jílek, 2001) a sú závislé od stupnice merania. Veľmi často používaná miera, ktorou je Giniho index (Giniho koeficient), nevyhovuje všetkým princípom axiomatického konceptu príjmovej nerovnosti.

2. Axiomatický koncept príjmovej nerovnosti

V tomto príspevku sa budeme venovať len tým mieram, ktoré vyhovujú tzv. axiomatickému konceptu. Miery, ktorými meriame nerovnomernosť rozdelenia, by mali vyhovovať požiadavkám, ktoré sú obsiahnuté v nasledujúcich piatich axiómach (Cowell, 1985).

Pigou-Daltonov princíp prenosu (citlivosti prenosu) znamená, že pri transfere príjmov od chudobných smerom k bohatým by sa mala hodnota miery príjmovej nerovnosti zvýšiť (nemala by klesnúť). Prenos príjmov od bohatých k chudobným by mal viesť ku zvýšeniu miery nerovnosti (hodnota by nemala klesnúť).

Predpokladajme, že $\mathbf{y}$ je vektor príjmov $y_1, y_2, ... y_n$, kde n vyjadruje počet jednotiek $o_1, o_2, ... o_n$ danej populácie (jednotlivcov, domácností,...) a $\mathbf{y}'$ je vektor príjmov, ktorý vznikol

transferom príjmu δ medzi objektmi s príjmami y_j a y_i vektora príjmov $\mathbf{y}$. Ak platí $y_i > y_j$, pričom $y_i + \delta > y_j - \delta$, podmienka je splnená ak pre miery $I(\mathbf{y}')$ a $I(\mathbf{y})$ platí nerovnosť $I(\mathbf{y}') \geq I(\mathbf{y})$.

Princíp symetrie (anonymity) je definovaný ako schopnosť miery zachovať si svoju veľkosť, ak dôjde k výmene príjmov medzi dvojicou jednotlivcov alebo domácností. To znamená, že veľkosť miery nie je ovplyvnená inými charakteristikami jednotlivcov alebo domácností, okrem tých charakteristík, ktorých nerovnomernosť rozdelenia je meraná. Pre všetky permutácie $\mathbf{y}$ a $\mathbf{y}'$ platí: $I(\mathbf{y}') = I(\mathbf{y})$.

Princíp nezávislosti priemeru vyjadruje takúto vlastnosť miery: ak sa všetky príjmy zvýšia λ - násobne, kde λ je ľubovoľná konštanta, veľkosť miery sa nezmení: $I(\mathbf{y}) = I(\mathbf{y} \cdot \lambda)$

Princíp populačnej homogénnosti vyžaduje od miery nerovnosti, aby sa nemenila, ak sa zmení veľkosť populácie. To znamená, že veľkosť miery sa nezmení, ak dôjde k zlúčeniu identických rozdelení. Pre ľubovoľnú konštantu $\lambda > 0$ platí: $I(\mathbf{y}) = I(\mathbf{y}[\lambda])$, kde $\mathbf{y}[\lambda]$ vyjadruje λ - násobnú replikáciu populácie.

Princíp rozložiteľnosti. Ak je splnená požiadavka rozložiteľnosti, môžeme mieru nerovnosti pre celú populáciu vypočítať aj na základe jej hodnôt pre podmnožiny, na ktoré je možné celú populáciu rozložiť bezo zvyšku. To znamená, že ak sa miera zvyšuje na podmnožinách celkovej populácie, možno očakávať jej zvýšenie na celej populačnej množine: $I_{TOTAL} = I_{BETWEEN} + I_{WITHIN}$, kde I_{TOTAL} je hodnota miery na celej populácii, $I_{BETWEEN}$ je medziskupinová hodnota miery a I_{WITHIN} je vnútroskupinová hodnota miery.

Všetky uvedené vlastnosti majú miery zo skupiny generalizovanej entropie a Atkinsonov index. Veľmi často používaný Giniho index nespĺňa podmienku rozložiteľnosti.

3. Miery generalizovanej entropie

Miery zo skupiny generalizovanej entropie majú všeobecný tvar:

$$GE(\alpha) = \frac{1}{\alpha^2 - \alpha}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\bar{y}}\right)^{\alpha} - 1\right] \tag{1}$$

kde y_i je príjem *i*-teho objektu, $(i = 1,2,..n)$, $\bar{y}$ je priemerná hodnota príjmu.

$GE(\alpha)$ nadobúda minimálnu hodnotu 0 v prípade príjmovej rovnosti, s rastom jej hodnôt sa zvyšuje príjmová nerovnosť. Koeficient α vo vzťahu (1) je váhou vzdialeností medzi príjmami v rôznych častiach príjmového rozdelenia a najčastejšie nadobúda hodnoty 0,1 a 2.

Pre hodnotou $\alpha = 0$ dostávame *Theilov L index*:

$$Theil\ L = GE(0) = \lim_{\alpha \to 0} \frac{1}{\alpha^2 - \alpha}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\bar{y}}\right)^{\alpha} - 1\right] = \frac{1}{n}\sum_{i=1}^{n}\log\frac{\bar{y}}{y_i} \tag{2}$$

Pre hodnotu $\alpha = 1$ dostávame *Theilov T index*:

$$Theil\ T = GE(1) = \lim_{\alpha \to 1} \frac{1}{\alpha^2 - \alpha} \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\bar{y}} \right)^{\alpha} - 1 \right] = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{\bar{y}} \log \frac{y_i}{\bar{y}} \tag{3}$$

4. Atkinsonov index nerovnosti

Ďalšiu skupinu mier predstavujú miery Atkinsonovho indexu nerovnosti:

$$A_{\varepsilon} = 1 - \left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i}{\bar{y}} \right)^{1-\varepsilon} \right]^{\frac{1}{1-\varepsilon}}, \ \varepsilon \neq 1 \tag{4}$$

$$A_{\varepsilon} = 1 - \frac{\prod_{i=1}^{n} y_i^{\frac{1}{n}}}{\bar{y}}, \varepsilon = 1 \tag{5}$$

Atkinsonov index nerovnosti je založený na „spravodlivom priemernom príjme“ $y_{\varepsilon} = \left[\frac{1}{n} \sum_{i=1}^{n} (y_i)^{1-\varepsilon} \right]^{\frac{1}{1-\varepsilon}}$. „Spravodlivý priemerný príjem“ je príjmom, ktorý ak je rovnomerne rozdelený v populácii, zabezpečí jej rovnakú úroveň blahobytu ako existujúce rozdelenie príjmov.

Vo vzťahoch (4) a (5) vystupuje parameter ε, ktorý sa nazýva parameter averzie voči rovnosti. Jeho nulová hodnota indikuje nezáujem spoločnosti o rozdelenie príjmov. Opakom je jeho extrémna hodnota ∞, ktorá je prejavom záujmu spoločnosti len o jednotlivca s najnižšou hodnotou príjmu.

Atkinsonov index nerovnosti nadobúda minimálnu hodnotu 0 v prípade absolútnej príjmovej rovnosti a maximálnu hodnotu 1 v prípade absolútnej príjmovej nerovnosti.

V tabuľke 1 sú na ilustráciu uvedené výpočty mier generalizovanej entropie a Atkinsonovho indexu pre rôzne hodnoty parametra averzie.

Tabuľka 2: Výpočet mier príjmovej nerovnosti

Objekty										Výpočty	Miera
10	15	20	25	40	20	30	35	45	90	y_i	
0,30	0,45	0,61	0,76	1,21	0,61	0,91	1,06	1,36	2,73	$y_i/\bar{y}$	
-0,52	-0,34	-0,22	-0,12	0,08	-0,22	-0,04	0,03	0,13	0,44	$\ln(y_i/\bar{y})$	
-0,16	-0,16	-0,13	-0,09	0,10	-0,13	-0,04	0,03	0,18	1,19	$(y_i/\bar{y})\ln(y_i/\bar{y})$	Theil T=0,080
0,52	0,34	0,22	0,12	-0,08	0,22	0,04	-0,03	-0,13	-0,44	$\ln(\bar{y}/y_i)$	TheilL=0,078
0,55	0,67	0,78	0,87	1,10	0,78	0,95	1,03	1,17	1,65	$(y_i/\bar{y})^{1/2}$	$A_{0,5}$ =0,087
1,26	1,31	1,35	1,38	1,45	1,35	1,41	1,43	1,46	1,57	$(y_i)^{1/n}$	A_1 = 0,164
3,30	2,20	1,65	1,32	0,83	1,65	1,10	0,94	0,73	0,37	$(y_i/\bar{y})^{-1}$	A_2 = 0,290

Zdroj: www.worldbank.org/poverty/inequal/index.htm

5. Rozklad mier nerovnosti

Ak je celá populácia rozdelená do niekoľkých populačných podmnožín, môžeme obidve vyššie uvedené miery I vyjadriť ako súčet tzv. vnútroskupinovej nerovnosti (*within-group*) I_W a medziskupinovej nerovnosti (*between group*) I_B.

V tomto príspevku uvedieme rozklad generalizovanej entropie. Vnútroskupinovú zložku nerovnosti určíme na základe vzťahu:

$$I_W = \sum_{j=1}^{k} w_j GE(\alpha)_j \text{, pričom } w_j = v_j^{\alpha} f_j^{1-\alpha} \tag{6}$$

kde f_j je podiel populácie *j*-tej podmožiny, v_j je podiel príjmu *j*-tej podmožiny a $GE(\alpha)_j$ je miera nerovnosti v *j*-tej podmnožine (j =1, 2, ...k).

Medziskupinovú zložku nerovnosti vypočítame takto:

$$I_B = \frac{1}{\alpha^2 - \alpha}\left[\sum_{j=1}^{k} f_j \left(\frac{\bar{y}_j}{\bar{y}}\right)^{\alpha} - 1\right] \tag{7}$$

Podiel I_B / I potom vyjadruje podiel regionálnych rozdielov na celkovej nerovnomernosti rozdelenia príjmov.

Theilovo L môžeme rozložiť na vnútroskupinovú a medziskupinovú nerovnosť:

$$Theil\ L = \sum_{j=1}^{k}\left(\frac{n_j}{n}\right) L_j + \sum_{j=1}^{k} \frac{n_j}{n} \log \frac{n_j / n}{y_j / y} \tag{8}$$

kde y_j je príjem a n_j početnosť *j*-tej podmnožiny, *y* je príjem, *n* je početnosť celej populácie a L_j je Theilovo L na *j*-tej podmnožine.

Rozklad *Theilovho T* je potom:

$$Theil\ T = \sum_{j=1}^{k}\left(\frac{y_j}{y}\right) T_j + \sum_{j=1}^{k}\left(\frac{y_j}{y}\right) \log \frac{y_j / y}{n_j / n} \tag{9}$$

Rozklad Theilovho indexu umožňuje identifikovať vplyv regionálnych nerovností na rôznych úrovniach regionálneho členenia (podmnožinách populácie) na príjmovú nerovnosť celej krajiny (celej populácie).

6. Záver

Pri meraní príjmovej nerovnosti sa najčastejšie používa Lorenzova krivka a Giniho koeficient koncentrácie. Existuje skupina mier, ktoré umožňujú analyzovať príjmovú nerovnosť populácie vzhľadom na existujúce disparity jej podmnožín, ktoré sú vytvorené na základe istých charakteristík jej subjektov. V tomto príspevku sme načrtli možnosti použitia Theilovho a Atkinsonovho indexu. Uvádzame spôsoby ich výpočtu a jeden z možných spôsobov rozkladu miery generalizovanej entropie.

7. Literatúra

[1]ATKINSON, A. B., 1970. On the Measurement of Inequality. In: Journal of Economic Theory, 1970, č. 2, s. 244-263.

[2]BARTOŠOVÁ, J. 2004. Statistický model příjmových rozdělení. In: Acta Universitatis Bohemiae Meridionales, vědecký časopis pro ekonomiku, řízení a obchod. 7(2). České Budějovice: Jihočeská univerzita, 2004, s. 39-46, ISSN 1212-3285.

[3]BARTOŠOVÁ, J. - STANKOVIČOVÁ, I. 2009: Diferenciace příjmú a chudoba v českých a slovenských domácnostech In: MSED 2009. Sborník přispěvků. - Praha : VŠE, 2009. ISBN 978-80-86175-66-9, elektronický dokument [CD-ROM]

[4]BOURGUIGNON, F. 1979. Decomposable Income Inequality Measures. In: Econometrica, 1979, 47, č. 4, s 901-920.

[5]COWELL, F.A. 1985. Measures of Distributional Change: An Axiomatic Approach. In: Review of Economic Studies. 52, 1985, s. 135-151.

[6]COWELL, F.A. – S.P. JENKINS. 1995. How Much Inequality can we Explain? A Methodology and an Application to the USA. In: Economic Journal, 105, 1995, s. 421-430.

[7]JÍLEK, J. a kol. 2001. Nástin sociáněhospodářské statistiky. Praha: Vydavatelstvo VŠE, 2001. 246 s. ISBN 80-245-0214-3.

[8]LAPÁČEK, M. Ekvivalenční stupnice a příjmová nerovnost. http://nf.vse.cz/download/veda/workshops/inequality.pdf

[9]LITCHFIELD, J. A. Inequality: Methods and Tools. www.worldbank.org/poverty/inequal/index.htm

[10] MUSSARD, S. et. al. 2003. Decomposition of Gini and the generalized entropy inequality measures. In: Economics Bulletin, 2003, Vol. 4, č. 7, s. 1–6

[11] SIPKOVÁ, Ľ. 2004. Prehľad teoretických východísk merania príjmovej nerovnosti. In Slovenská štatistika a demografia. Bratislava: Štatistický úrad Slovenskej republiky, 2004, roč. 14, č. 3, s. 36-49, ISSN 1210-1095.

[12] VOJTKOVÁ, M. 2010. The evaluation of regional employment disparities. In Data analysis methods in economics research, nr.11: studia i prace Uniwersytetu Ekonomicznego w Krakowie. - Kraków : Universytet Ekonomiczny w Krakowie, 2010. s. 125-136, ISBN 978-83-7252-478-2.

[13] ŽELINSKÝ, T. 2010. Nerovnosť rozdeľovania príjmov v krajoch Slovenskej republiky. In: Slovenská štatistika a demografia. č. 1, 2010, s.

Adresa autora:

Viera Labudová, RNDr., PhD.
FHI EU v Bratislave
Dolnozemská 1
852 35 Bratislava
viera.labudova@euba.sk

Tento článok vznikol v rámci projektu VEGA č. ***1/0437/08*** *Kvantitatívne metódy v stratégii šesť sigma.*

Neparametrická regresia
Nonparametric regression

Lenka Lašová, Veronika Valenčáková

Abstract: Contribution deals with nonparametric approach to the regression model. We describe main principles of this type of regression and focus on nonparametric regression based on kernel estimator. We illustrate this method on an example and we point out the situations when this approach is appropriate.

Key words: regression model, nonparametric regression, kernel estimator

Kľúčové slová: regresný model, neparametrická regresia, kernel-estimátor

JEL classification: C14

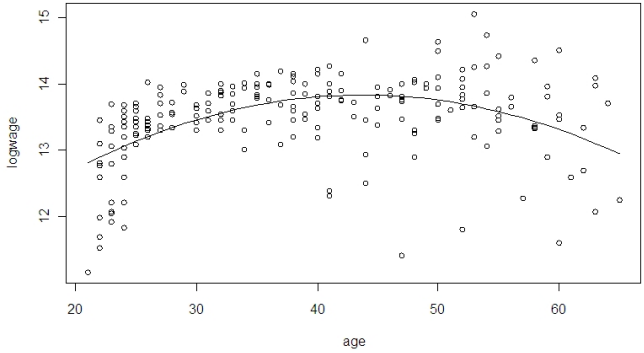
1. Úvod

Vychádzame z jednoduchého regresného problému. Teda máme k dispozícii hodnoty $y_1, y_2, \ldots, y_n$ namerané pre fixné $x_1, x_2, \ldots, x_n$ a predpokladáme, že platí model

$$y_i = f(x_i) + \varepsilon_i, \tag{1}$$

kde ε_i sú nezávislé rovnako rozdelené náhodné premenné s nulovou strednou hodnotou a neznámou varianciou σ^2. Naším cieľom je čo najlepšie odhadnúť neznámu funkciu *f*.

V prípade klasického parametrického prístupu postupujeme tak, že najprv zvolíme parametrickú triedu funkcií, z ktorej budeme vyberať odhad funkcie *f*. Keďže môžeme zvoliť akúkoľvek parametrickú triedu funkcií, tento prístup je pomerne flexibilný. Parametrický prístup má viaceré výhody: je efektívny v prípade, že náš model je korektný (teda sme dobre zvolili triedu parametrických funkcií), jeho výstupom presný predpis odhadovanej funkcie a jej parametre sa dajú obvykle dobre interpretovať. Teda parametrickú regresiu by sme mali použiť vždy, keď vopred poznáme typ odhadovanej funkcie. Problém však môže nastať, ak sme zle špecifikovali model, teda sme zle zvolili triedu hustôt. Nami odhadnutá funkcia môže byť v tomto prípade značne vychýlená od skutočnosti. Prípad takto zle zvolenej triedy ilustrujeme na obrázku 1.



Obrázok 2: Vstupné dáta a odhadnutá parametrická regresná krivka

Ide o dáta, v ktorých skúmame závislosť (logaritmovaného) príjmu od veku. Použili sme triedu kvadratických funkcií a odhadli sme funkciu *f* v rámci tejto triedy. Získaný model sme podrobili testu vzhľadom na správnu špecifikáciu modelu [2], ktorým sme zamietli hypotézu o správnej špecifikácii modelu (na hladine menšej ako 0,001).

V ďalšom texte stručne popíšeme neparametrický prístup k regresii a ukážeme jeho využitie na spomínanom príklade.

2. Odhad hustoty pomocou kernel-estimátora

Najprv si povieme ako neparametricky odhadovať hustotu z náhodného výberu. Na odhad hustoty $f(\mathbf{x})$ možno použiť viacero estimátorov. Medzi často používané neparametrické estimátory patrí tzv. kernel-estimátor. Pre jednoduchosť popíšeme odhad funkcie hustoty použitím tohto estimátora v prípade jednorozmernej náhodnej premennej.

Nech X je náhodná premenná s hustotou $f(x)$ a $x_1, x_2, \ldots, x_n$ je jej náhodný výber. Potom odhad funkcie *f* v bode x_0 je

$$\hat{f}(x_0, x_1^n) = \frac{\sum_{i=1}^{n} K\left(\frac{x_0 - x_i}{h}\right)}{nh}, \tag{2}$$

kde funkcia K je nazývaná kernel-estimátor a spĺňa nasledujúce požiadavky: je symetrická okolo nuly, po častiach spojitá, nezáporná a $\int_R K(x)dx = 1$.

Zvoľme za *K* funkciu definovanú predpisom:

$$K(\mu) = \frac{I_{(-1,1)}(\mu)}{2}, \tag{3}$$

kde $I_{(-1,1)}(\mu)$ je funkcia, ktorá bodom v intervale $(-1, 1)$ priradí hodnotu 1 a ostatným bodom priradí 0. Odhad celej funkcie *f* si potom môžeme ľahko predstaviť ako kĺzavý priemer nameraných hodnôt s dĺžkou intervalu 2*h*.

3. Základný popis metódy

V tejto časti popíšeme neparametrickú regresiu s využitím kernel-estimátora. Majme náhodný vektor $(Y, \mathbf{X})$ so združenou hustotou $f_{Y,\mathbf{X}}(y, \mathbf{x})$, kde $\mathbf{X}$ je *k*-rozmerný vektor. Označme skutočnú regresnú funkciu $g(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ a predpokladajme, že $\mathrm{Cov}(Y|\mathbf{X}) = \sigma^2\mathbf{I}$, $E(\varepsilon|\mathbf{X}) = 0$, $\mathrm{Cov}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}$. Pri tomto type regresie použijeme vzťah:

$$E(Y|\boldsymbol{X}) = \int_{-\infty}^{\infty} y \frac{f(y,x)}{f(x)} dy, \tag{4}$$

pričom pri výpočte odhadu funkcie $g(\mathbf{x})$ dosadíme za $f(y, \mathbf{x})$, $f(\mathbf{x})$ práve odhady vypočítané pomocou kernel-estimátora. Veľkou výhodou neparametrického prístupu je, že hľadá funkciu *f* z triedy hladkých funkcií bez toho, aby bolo vopred potrebné danú funkciu špecifikovať. Teda rozsah potencionálnych funkcií, ktorými môžeme odhadnúť neznámu funkciu *f*, je oveľa širší než v parametrickom prípade, čiže tento prístup je oveľa flexibilnejší. Aj keď je rozumné požadovať od hľadanej funkcie niektoré vlastnosti (napríklad určitý stupeň hladkosti a spojitosť), sú tieto požiadavky neporovnateľne menej limitujúce ako definovanie presného typu funkcie v prípade parametrickej regresie.

Neparametrická regresia je menej náchylná k veľkým chybám, keďže kladie menej predpokladov na odhadovanú funkciu *f*. Má však aj nevýhody - asi najväčšou z nich je, že výstupom metódy nie je priame vyjadrenie vzťahu prediktorov a závislých premenných, ale zväčša je to len grafický výstup a prípadne hodnoty v bodoch, ale nie priamy predpis funkcie. S tým súvisí aj problém interpretácie výsledkov.

Po získaní odhadov hustôt $f(y,\mathbf{x})$, $f(\mathbf{x})$ tieto odhady dosadíme do vzorca (4), čím dostávame Nadaraya Watsonov regresný odhad funkcie *g*:

$$\hat{g}(\mathbf{x}_0,\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{\mathbf{x}_0 - \mathbf{x}'_{i\bullet}}{\mathbf{h}}\right)}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_0 - \mathbf{x}'_{i\bullet}}{\mathbf{h}}\right)} = \sum_{i=1}^{n} w_i(\mathbf{x}_0) y_i \,. \tag{5}$$

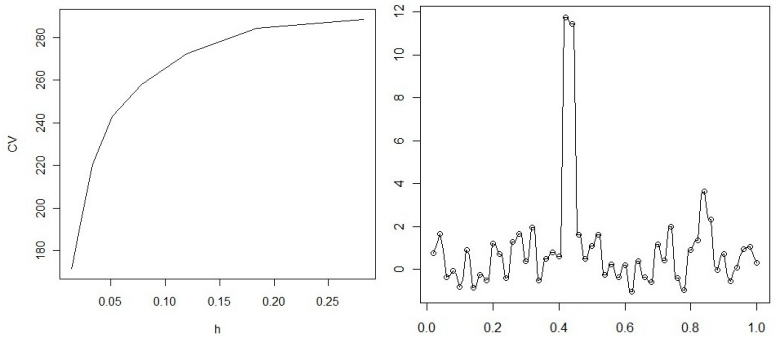
Ak by sme za kernel-estimátor zobrali funkciu definovanú rovnosťou (4), tak ako odhad funkcie *g* v bode $\mathbf{x}_0$ dostaneme priemer tých hodnôt *y*, pre ktoré $\mathbf{x}$ leží v intervale $(\mathbf{x_0} - \mathbf{h}, \mathbf{x_0} + \mathbf{h})$. Pre akýkoľvek iný kernel-estimátor je hodnota odhadu funkcie *g* v bode $\mathbf{x}_0$ vážený priemer takýchto *y*. Inými slovami, hodnotu v bode $\mathbf{x}_0$ odhadujeme konštantnou funkciou, ktorú vždy stanovíme len na základe nameraných hodnôt, ktoré majú svoj vzor v intervale dĺžky 2$\mathbf{h}$ so stredom $\mathbf{x}_0$. Toto vedie k prirodzenému zovšeobecneniu tejto metódy, a to použiť na odhadovanie namiesto konštantných funkcií akýkoľvek polynóm rádu *r*. Takýmto spôsobom sa dostávame k semiparametrickému modelu známemu pod názvom lokálna polynomická regresia a Nadaraya Watsonov odhad môžeme chápať ako lokálnu polynomickú regresiu nultého rádu.

Dôležitou otázkou zostáva vhodná voľba kernel-estimátora a tiež dĺžky intervalu *h*. Vo všeobecnosti výber kernel-estimátora nemá veľký vplyv na vlastnosti Nadaraya Watsonovho odhadu, ale je vhodné voliť hladké, kompaktné funkcie (my sme používali Gaussovský kernel-estimátor). Dĺžka intervalu *h* slúži ako vyhladzujúci parameter, teda čím je dĺžka intervalu väčšia, tým je krivka hladšia. Ak je však interval príliš dlhý, nemusí zachytiť dôležité výkyvy v dátach. Obvykle vyskúšame viacero možností a výsledky graficky porovnáme. Subjektívny výber *h* však vyžaduje veľa času, preto sú často užitočné automatické metódy na voľbu dĺžky intervalu *h*.

Keď disponujeme informáciou o tom, ako by mal vyzerať hľadaný regresný model, mali by sme dĺžku intervalu *h* vždy určiť subjektívne. V opačnom prípade (keď nemáme dostatok informácií o modeli) je vhodné použiť na prvotný prieskum nejaký zautomatizovaný výber dĺžky intervalu *h*. Jednou z možných metód je napríklad cross-validácia, kedy hľadáme hodnotu *h*, ktorá minimalizuje funkciu *CV*:

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{g}_{h(i)}(x_i)\right)^2 \,. \tag{6}$$

Táto metóda výberu *h* však nie je univerzálna – na ilustráciu uvedieme príklad, kedy uvedená metóda zlyhá. Na obrázku 2 vľavo je graficky znázornená funkcia *CV* . Vidno, že funkcia nadobúda minimum pre *h* idúce k nule. Na obrázku 2 vpravo je výsledná odhadnutá funkcia pre hodnotu *h*=0,009, ktorá je však úplne nevhodná, pretože kopíruje dáta.

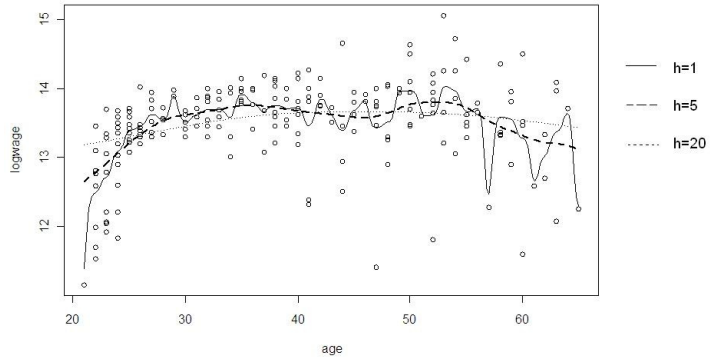


Obrázok 2: Funkcia CV (vľavo) a výsledná regresná funkcia pre h=0,009 (vpravo)

Ak sme vybrali optimálnu dĺžku intervalu *h*, tak priemerná chyba štvorcov, ktorou zvykneme merať kvalitu modelu aj v prípade parametrickej regresie, konverguje k nule rýchlosťou $O(n^{-4/5})$. Teda môžeme skonštatovať, že v porovnaní s dobre špecifikovaným parametrickým modelom (ktorý konverguje rýchlosťou $O(n^{-1})$) je na tom náš model horšie. Ale veľkou výhodou neparametrického modelovania je ochrana proti zlej špecifikácii modelu. Ak sme pri parametrickom prístupe zvolili zlú triedu funkcií, tak priemerná chyba štvorcov konverguje rýchlosťou O(1), teda takýto odhad sa pre zväčšujúci sa výber vôbec nezlepšuje.

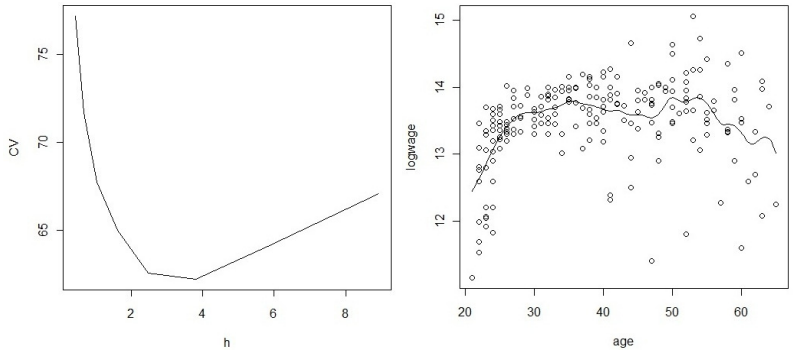
4. Použitie neparametrickej regresie

Použijeme neparametrickú regresiu na vstupné dáta zobrazené na obrázku 1, kde parametrický model zlyhal. Použijeme tri rôzne dĺžky intervalu *h*, aby sme ilustrovali rozdielne regresné krivky (obrázok 3). V prvom prípade sme použili *h*=20, čo je v porovnaní s rozsahom dát pomerne vysoká hodnota. Na grafe vidíme, že takáto funkcia je síce hladká ale veľmi sa podobá regresnej funkcii z parametrického modelu a nie je schopná dobre zachytiť výkyvy v dátach. Keď zvolíme *h*=1, vidíme vďaka príliš nízkej hodnote opačný extrémny prípad – funkcia je až príliš zviazaná s dátami a absentuje vyhladenie. Nakoniec sme zvolili *h*=5, v tomto prípade je výsledná funkcia dostatočne hladká a ako si môžeme všimnúť, poukazuje na jeden dôležitý výkyv v dátach pre respondentov medzi 40. - 50. rokom života.



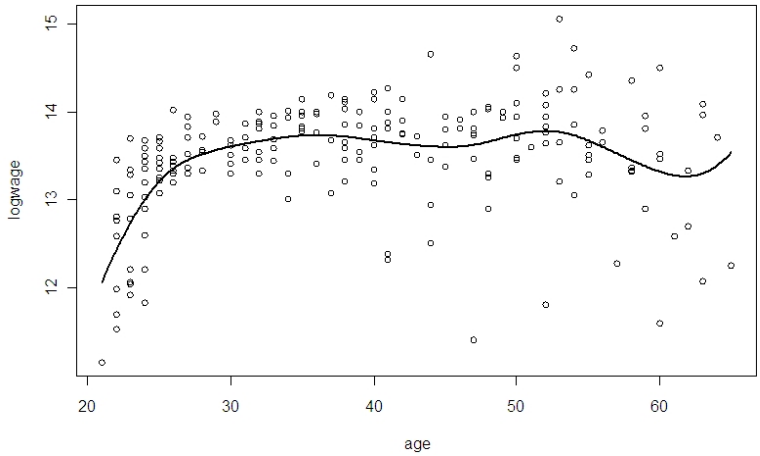
Obrázok 3: Neparametrické regresné funkcie pre rôzne hodnoty h

Voľba dĺžky intervalu *h* pomocou cross-validácie je znázornená na obrázku 4. Vidíme, že automatický výber optimálnej hodnoty *h* pre vytvorenie regresnej funkcie síce nezanedbáva výkyvy v dátach, ale výsledná funkcia nie je príliš hladká a teda nemusí dostatočne vyhovovať našim požiadavkam. Aj v tomto prípade je teda vhodné pozrieť sa, čo nám metóda cross-validácie radí ale nie je vhodné jej výber *h* použiť pre konečný model.



Obrázok 4: Neparametrická regresia s využitím cross-validácie: funkcia CV (vľavo), výsledná regresná funkcia (vpravo)

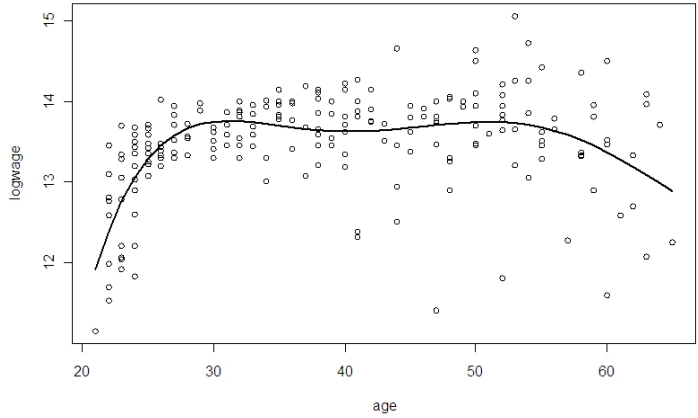
Na obrázku 5 je znázornený odhad regresnej funkcie pomocou lokálnej polynomickej regresie s dĺžkou intervalu *h*=2,5. Aj na tomto obrázku môžeme vidieť evidentný výkyv v dátach medzi 40. - 50. rokom.



Obrázok 5: Lokálna polynomická regresia – výsledná regresná funkcia pre h=2,5

Na základe regresných funkcií získaných neparametrickou regresiou sme v dátach identifikovali výkyv, ktorý parametrická regresia neodhalila. Z grafického znázornenia regresných funkcií tiež môžeme približne identifikovať triedu funkcií, ktorá je vhodná na modelovanie daných dát. Avšak ako sme už spomenuli v úvode, pomocou tejto metódy

nezískame priamy predpis odhadnutej regresnej funkcie. Je preto vhodné využiť získané informácie pre vytvorenie modelu pomocou parametrickej regresie, aby sme získali aj predpis odhadu funkcie *g*. Na základe týchto poznatkov zvolíme za parametrickú triedu polynomických funkcií 5. stupňa. Na obrázku 6 je znázornená výsledná odhadnutá funkcia *g*. Hypotézu o správnej špecifikácii tohto regresného modelu na základe testu korektnosti modelu nezamietame (p-hodnota testu je 0,68672).



Obrázok 6: Výsledná odhadnutá regresná funkcia získaná parametrickou regresiou

5. Literatúra

[1]FARAWAY, J.J. 2006. Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman & Hall/CRC 2006. 301 s. ISBN 0-203-49228-5

[2]HSIAO, C. – LI, Q. – RACINE, J. 2004. A consistent model specification test with mixed categorical and continuous data, Unpublished manuscript, Department of Economics, Texas A & M University.

[3]LI, Q. – RACINE, J. 2007. Nonparametric Econometrics: Theory and Practice, Princeton U. Press 2007.

[4]RACINE, J. 1997. Consistent significance testing for nonparametric regression, J. Business Economics and Stat, 15(3).

Adresa autora (-ov):

Lenka Lašová, RNDr., PhD.
Inštitút matematiky a informatiky ÚVVUMB
Cesta na amfiteáter 1
974 01 Banská Bystrica
lasova@fpv.umb.sk

Veronika Valenčáková, RNDr., PhD.
Katedra matematiky FPV UMB
Tajovského 40
974 01 Banská Bystrica
valencak@fpv.umb.sk

Tento príspevok vznikol vďaka podpore v rámci OP Výskum a vývoj pre projekt: CaKS - Centrum excelentnosti informatických vied a znalostných systémov (ITMS: 26220120007), spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja a vďaka podpore UGA pre projekt I-10-005-07.

Intervaly spoľahlivosti pre poistné rezervy
Confidence intervals for claim reserves

Bohdan Linda, Jana Kubanová, Pavla Jindrová

Abstract: The paper shows solution of the problem of confidence interval construction for predicted cumulative Chain ladder claims.

Abstrakt: V príspevku je ukázané riešenie problému konštrukcie intervalu spoľahlivosti pre rezervy na poistné plnenia.

Key words: insurance benefit, claim reserve, development factor, Chain ladder method, confidence interval

Kľúčové slová: poistné plnenie, poistná, vývojový faktor, metóda Chain ladder, intervalové odhady

JEL classification: C, G

1. Introduction

When the insurance claim reserves are planned the insurance companies result from the data about the insurance benefit in preceding years. Many methods are used to determine the amount of these reserves, one of the most frequent methods is the Chain ladder (CL) method (Pacáková 2004), (Mack 1993).

The results of the basic Chain ladder method can be influenced by extreme damages that are not common in each year and therefore the resulting values needn't to be reliable enough.

This lack could be removed, if we know the probability distribution of the random variable, expressing the insurance benefit. To be able to estimate this probability distribution by the exact statistical method we usually don't have sufficient number of observations. That is why many different approximative methods are used to calculate these claim reserves.

Recently the bootstrap methods are actually applied. Many of mentioned methods are described in (Wüthrich, Merz 2008).

Results of these methods are the point estimates of the claim reserves. The interval estimates are often preferred to point estimates, because they provide more information about the estimated values. The goal of this paper is to present the confidence interval construction for claim reserves and for the development factor κ by the help of the bootstrap method.

2. The basic process of the confidence interval construction

The standard form of the confidence interval in the case that the parameter's estimate $\hat{\Theta}$ of the parameter Θ is the function of the sum of the random variables is

$$I_{1-2\alpha} = \langle \hat{\Theta} - t_{1-\alpha,\, n-1} \cdot SE;\ \hat{\Theta} - t_{\alpha,\, n-1} \cdot SE \rangle,$$

where $t_{1-\alpha,n-1}$ and $t_{\alpha,n-1}$ are 1-α-quantile and α-quantile of the Student's probability distribution with $n - 1$ degrees of freedom and SE is the standard error. The form of this interval is derived from the fact that the random variable $\frac{\hat{\Theta} - \Theta}{SE}\sqrt{n}$ can be approximated by the Student's probability distribution.

The above stated formula generally doesn't provide the correct interval estimation of the parameter Θ in the case of the small range of the sample that is from different than normal distribution. The technique of the construction of the confidence interval based on bootstrap can be applied for this reason and the confidence intervals without any constraining

presumptions can be obtained. The method is based on bootstrapping of probability distribution of the statistics $\frac{\hat{\Theta}-\Theta}{SE}\sqrt{n}$. Such interval is called bootstrap *t*-interval. This method is described in a detailed way in (Davison, Hinkley 1997). The stated method can be considered as generalization of the commonly known Student's method. It can be applied when characteristics of position are estimated, also characteristics like sample average, median, cut average or sample quantile. The bootstrap method, in the simplest form like this described above, is not proper for solution of the general problems. That is why it is necessary to search more reliable procedures. One of them is confidence interval based on bootstrap quantile. (Kubanová 2003).

3. The bootstrap quantile interval

We assume that R bootstrap samples $\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*R}$ from distribution $G_n(x, \hat{F})$ are generated. The bootstrap replications $\hat{\Theta}^*(i) = \hat{\Theta}(\boldsymbol{x}^{*i})$ for each sample are calculated. Let $\hat{G}_R$ be the distribution function of the variable $\hat{\Theta}^*$.

1 - 2α quantile interval is defined by the help of α and 1 - α quantile of $\hat{G}_R$:

$$\langle \hat{\Theta}_{low} ; \hat{\Theta}_{up} \rangle = \langle \hat{G}_R^{-1}(\alpha) ; \hat{G}_R^{-1}(1 - \alpha) \rangle. \qquad (1)$$

$\hat{G}_R^{-1}(\alpha) = \hat{\Theta}_{\alpha}^{*}$ is the α-quantile of bootstrap distribution according to the definition. The quantile interval can as well written in the form

$$\langle \hat{\Theta}_{low} ; \hat{\Theta}_{up} \rangle = \langle \hat{\Theta}_{\alpha}^{*} ; \hat{\Theta}_{1-\alpha}^{*} \rangle . \qquad (2)$$

The expressions (1) and (2) present the ideal situation, when infinitely bootstrap simulation are done, it means the situation when $R \to \infty$. In reality, it is possible to realize only finite number of simulations.

Let $\hat{\Theta}_{R,\alpha}^{*}$ be α empirical quantile of the values $\hat{\Theta}^*$, which is the $R \cdot \alpha$-th ordered value of R replications of the variable $\hat{\Theta}^*$. Similarly $\hat{\Theta}_{R,1-\alpha}^{*}$ is the (1 - α) empirical quantile.

The approximation of the 1 - 2α quantile interval is then

$$\langle \hat{\Theta}_{low}; \hat{\Theta}_{up} \rangle \sim \langle \hat{\Theta}_{R,\alpha}^{*} ; \hat{\Theta}_{R,1-\alpha}^{*} \rangle. \qquad (3)$$

4. The interval construction

To demonstrate the quantile bootstrap confidence interval, the data published by (Taylor, Ash 1983) were used. The presented method can be summarized in the following steps:

a) Table of cumulative observed data creation, see the claims development (upper) triangle in the table 2. This data express the cumulative claims in individual accident year and development year.

b) The development factor κ-estimate: the factor k_j is estimated by

$$k_j = \frac{\sum_{i=0}^{n-j-1} C_{i,j+1}}{\sum_{i=0}^{n-j-1} C_{i,j}} ; \; j = 0, 1, \ldots, n\text{-}1 \qquad (4)$$

c) The lower triangle matrix calculation: $C_{i,j+1} = k_j \cdot C_{i,j}$, (5)

d) Bootstrap claim reserving calculation based on AR(1) model use the more complicated theoretical model

$$C_{i,j+1} = \kappa_j \cdot C_{i,j} + \sigma_j \sqrt{C_{i,j}}\, \varepsilon_{i,j+1}, \tag{6}$$

to express the relation between $C_{i,j+1}$ and $C_{i,j}$.

e) Coefficient σ_j in the formula (6) is estimated by

$$s_j = \sqrt{\frac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - k_j \right)^2}\ ; \quad j = 0, 1, \ldots, n\text{-}1 \tag{7}$$

f) Residuals $\varepsilon_{i,j}$ are estimated by variable $e_{i,j}$ according to the formula (Wütrich, Merz, 2008):

$$e_{i,j+1} = \frac{\frac{C_{i,j+1}}{C_{i,j}} - k_j}{\sqrt{\frac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - k_j \right)^2}} \sqrt{C_{i,j}}\ .$$

g) The adjusted residuals $z_{i,j}$ instead of residuals $e_{i,j}$ are recommended to be used, which can be obtained according to the following formula:

$$z_{i,j+1} = \frac{\frac{C_{i,j+1}}{C_{i,j}} - k_j}{\sqrt{\frac{1}{n-j-1} \sum_{i=0}^{n-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - k_j \right)^2}} \sqrt{C_{i,j}} \Bigg/ \sqrt{1 - \frac{C_{i,j}}{\sum_{i=0}^{n-j} C_{i,j}}}\ . \tag{8}$$

h) **Bootstrap process.** The calculated values of the adjusted residuals are used in the bootstrap process. We assume that each value has the same probability and then the non-parametric bootstrap is used (Efron, Tibshirany 1993). The result of this process is the new upper triangle matrix, elements of which are indicated $z^*_{i,j}$.

g) By the help of these adjusted bootstrap residuals $z^*_{i,j}$ the new upper triangle matrix of the values $C^*_{i,j}$ is calculated by the following way:

For the first column we write: $C^*_{i,0} = C_{i,0}$.

The other values $C^*_{i,j}$ for $i + j < n$ are calculated according to the formula.

$$C^*_{i,j} = k_{j-1} C^*_{i,j-1} + s_{j-1} \sqrt{C^*_{i,j-1}}\ z^*_{i,j} \tag{9}$$

i) The bootstrap replications k_j^* of the development coefficient k_j are calculated

$$k^*_j = \frac{\sum_{i=0}^{n-j-1} C^*_{i,j+1}}{\sum_{i=1}^{I-j} C^*_{i,j}},\ j = 0, 1, \ldots, n\text{-}1 \tag{10}$$

j) The bootstrap values $C^*_{i,j}$ of the predicted cumulative bootstrap CL claims (lower triangle matrix) are calculated according to the formula

$$C^*_{i,j} = C_{i,n-i} \prod_{m=n-i}^{j-1} k^*_m . \qquad (11)$$

k) The bootstrap process from a) till i) is repeated *R*–times, after each repetition the vector of k_j^* , j = 0, 1, …, n-1, and lower triangle matrix of the predicted cumulative bootstrap CL claims $C^*_{i,j}$ are gained.

k) The limits of the bootstrap confidence interval are: $k^*_{R,\alpha}$ and $k^*_{R,1-\alpha}$, where $k^*_{R,\alpha}$ is the $R\cdot\alpha$-th ordered value of R replications of the variable k^* and similarly $k^*_{R,1-\alpha}$ is the $R\cdot(1-\alpha)$-th ordered value of R replications of the variable k^*.

Approximation of the 1 - 2α quantile interval $\langle k_{low}\,;\, k_{up}\rangle$ is then $\langle k^*_{R,\alpha}\,;\, k^*_{R,1-\alpha}\rangle$

5. Practical application of the method and results

The central point of interest in the Chain ladder method is developing factor k calculation. 90% confidence intervals for these factors k_j are stated on the table 1. The first row (bold) shows the point estimates of the developing factor for developing periods j and j+1.

Table 1: 90% confidence interval for development factors

R=1000	k_0	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8
	3.491	**1.747**	**1.457**	**1.174**	**1.104**	**1.086**	**1.054**	**1.077**	**1.018**
lower limit $k^*_{R,\alpha}$	3.129	1.659	1.376	1.131	1.063	1.050	1.044	1.058	1.018
upper limit $k^*_{R,1-\alpha}$	3.839	1.847	1.542	1.221	1.147	1.125	1.063	1.094	1.018

The figure 1 shows convergence of the development factor k_0^* when 1000 simulations were made and 1000 replication k_0^* calculated. We can state that after 500 replication the difference between originally calculated development factor k_0= 3,491 and bootstrap replication k_0^* is less than 0,005.

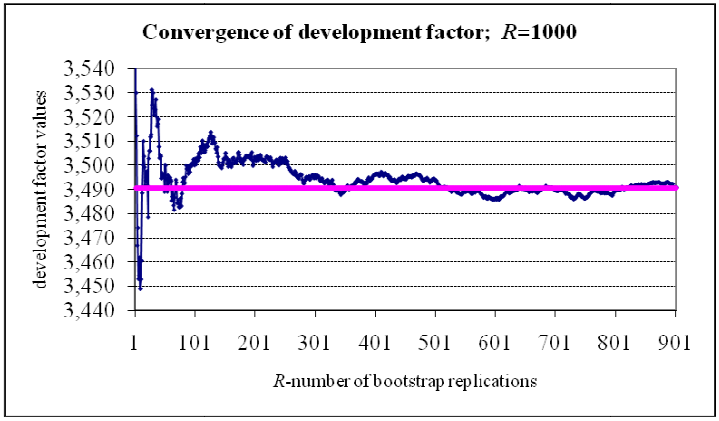


Figure 3: Convergence for development factor k_0^*, comparison with calculated development factor k_0= 3,491 (horizontal straight line)

The following table 2 presents the original claims development triangle (upper triangle) and three values in the lower triangle. The middle value represents the predicted cumulative CL claims, calculated by classical CL method, lower and upper values represent lower and upper limit of the 90% bootstrap confidence interval, obtained after 1000 bootstrap replications.

As can we see in the presented table 2, for example, that the length of the 90% confidence interval for $C_{7,6}$ creates 17% from forecasted value. It means that the real cumulated insurance benefit can differ from assumed within the framework of selected reliability 0,9 approximately about 8,5%. These 8,5% can create considerable financial sum at bigger volumes of insurance benefits. The insurance company should take this fact into account. When we would want to choose higher reliability, this percentage is even bigger.

6. Conclusion

The insurance companies should consider not only point estimates, but interval estimates as well at determination of the level of costs for claim reserves, as results from the part five of this paper. The fundamental problem is that it is necessary to know the probability distribution of the insurance benefit for the exact calculation of the confidence interval. This probability distribution can be with difficulty estimated with sufficient reliability with regard to a small number of data. For this reason it is necessary to use any approximative method when confidence intervals are calculated. Approximations based on central limit theorem are in this case relatively complicated. The article presents one application of the bootstrap method at confidence interval calculation; application of this method is relatively simple and it can be used on very general assumptions.

7. Literatúra

[1] EFRON, B, TIBSHIRANY, R. J. 1993. An Introduction to the Bootstrap. Chapman&Hall/CRC, 1993. ISBN 0-412-04231-2.
[2] HINKLEY, D. V, DAVISON, A.C.1997. Bootstrap methods and their application. Cambridge University Press1997. 582 s. ISBN 0-521-57391-2.
[3] KUBANOVÁ, J. 2003. Bootstrapové metody. Monografie. Univerzita Pardubice, 2003.120s. ISBN 80-7194-581-1
[4] LINDA, B., KUBANOVÁ, J., JINDROVÁ, P. 2010. Insurance reserves estimation by bootstrap. In: Proceedings of conference: Modelling, simulation and management of insured risks. Univerzita Pardubice 20101.
[5] MACK,T. 1993. Distribution-free calculation of the standard error of chain ladder reserve estimates. Astin Bullerin 23/2, s.213-225
[6] PACÁKOVÁ, V. 2004. Poistná štatistika. Iura Edition. Bratislava 2004. ISBN 80-8078-00-8.
[7] RUBLÍKOVÁ, E. *Analýza časových radov*. Iura Edition. Bratislava 2007. ISBN 978-80-8078-139-2.
[8] TAYLOR, G. C., ASHE, F. R. 1983. Second moments of estimates of outstanding claims. J. ECONOMETRICS 23, 37-61
[9] WÜTHRICH, M.V., MERZ, M. 2008. Stochastic Claims Reserving Methods in Insurance. John Wiley 2008. 424 s.ISBN 978-0-470-72346-3

Authors´ addresses:

Bohdan Linda,doc.RNDr.CSc.
Studentská 95
53210 Pardubice
bohdan.linda@upce.cz

Jana Kubanová,doc.PaedDr.CSc.
Studentská 95
53210 Pardubice
jana.kubanova@upce.cz

Mgr. Pavla Jindrová
Studentská 95
53210 Pardubice
pavla.jindrova@upce.cz

Table 2: 90% confidence interval for cumulative claims reserves

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0	357848	1124788	1735330	2218270	2745596	3319994	3466336	3606286	3833515	3901463
1	352118	1236139	2170033	3353322	3799067	4120063	4647867	4914039	5339085	**5433719** 5433719 **5433719**
2	290507	1292306	2218525	3235179	3985995	4132918	4628910	4909315	**5196236** 5285148 **5370601**	**5288338** 5378826 **5465793**
3	310608	1418858	2195047	3757447	4029929	4381982	4588268	**4790105** 4835458 **4877914**	**5108111** 5205637 **5301072**	**5198651** 5297906 **5395032**
4	443160	1136350	2128333	2897821	3402672	3873311	**4065215** 4207459 **4355925**	**4285262** 4434133 **4597033**	**4597928** 4773589 **4954707**	**4679425** 4858200 **5042527**
5	396132	1333217	2180715	2985752	3691712	**3923405** 4074999 **4233233**	**4199117** 4426546 **4671469**	**4425097** 4665023 **4929563**	**4748626** 5022155 **5300570**	**4832794** 5111171 **5394521**
6	440832	1288463	2419861	3483130	**3938235** 4088678 **4252261**	**4270918** 4513179 **4788181**	**4603113** 4902528 **5246295**	**4857304** 5166649 **5526383**	**5199827** 5562182 **5973396**	**5291992** 5660771 **6079273**
7	359480	1421128	2864498	**3940427** 4174756 **4415549**	**4578991** 4900545 **5221787**	**5002914** 5409337 **5821318**	**5391338** 5875997 **6391095**	**5674132** 6192562 **6735760**	**6089600** 6666635 **7271047**	**6197536** 6784799 **7399925**
8	376686	1363294	**2261025** 2382128 **2518111**	**3206700** 3471744 **3744930**	**3744319** 4075313 **4423418**	**4094217** 4498426 **4920496**	**4433872** 4886502 **5373563**	**4664225** 5149760 **5671035**	**4996437** 5544000 **6105410**	**5084997** 5642266 **6213626**
9	344014	**1076234** 1200818 **1320494**	**1870247** 2098228 **2363099**	**2672327** 3057984 **3473683**	**3132634** 3589620 **4086749**	**3455271** 3962307 **4541920**	**3717913** 4304132 **4959268**	**3903549** 4536015 **5223927**	**4200054** 4883270 **5619522**	**4274499** 4969825 **5719127**

Remark: *The article was elaborated with the support of the grant GAČR 402/09/1866, Modelling, simulations and management of insurance risk.*

Názory mladých v SR a ČR na niektoré javy súčasného populačného vývoja
Opinion of young people in SR and CR on some events of contemporary population evolution

Ján Luha, Dušana Zvarová

Abstract: In this article we present principal results from survey about opinion young people, university students, in Slovak republic and Czech Republic on selected events of contemporary population evolution.

Key words: population evolution, opinion of young people, comparison of results between SR and CR.

Kľúčové slová: populačný vývoj, názory mladých, porovnanie výsledkov medzi SR a ČR.

JEL Classification: C1, C12, C14, J10, J11.

1. Úvod

V príspevku uvádzame výsledky vlastného prieskumu názorov mladých ľudí, univerzitných študentov, v Slovenskej a Českej republike na niektoré javy v súčasnom populačnom vývoji. Prieskum bol projektovaný tak, aby boli získané názory mladých v SR a ČR študujúcich v hlavnom meste a v inom meste krajiny. V SR tým „iným" mestom bol Trenčín a v ČR Pardubice. Inšpiráciou a faktickým základom dotazníka, ktorý bol na prieskum použitý, bol výskum prof. Jozefa Mládka z roku 2004 (Mládek J., Širočková, J.,2004). S jeho osobným súhlasom boli využité formulácie v dotazníku pre aktuálny prieskum. Dotazník pre ČR mal modifikovaný iba znak kraj trvalého bydliska študenta. V článku uvádzame základné výsledky komparácie názorov mladých v SR a ČR.

2. Základné charakteristiky prieskumu

Na prieskume sa zúčastnili 422 univerzitných študentov, z toho bolo 262 z ČR a 160 zo SR. ČR je zastúpená študentmi VŠE Praha (FIS) v počte 163 a Univerzitou Pardubice v počte 99 študentov FSE. V SR boli získané odpovede 87 študentov LF UK a 23 študentov Pedagogickej fakulty UK v Bratislave a 50 študentov Univerzity A. Dubčeka v Trenčíne. Rozloženie podľa pohlavia v SR je 72,5% ženy a 27,5% muži a v ČR 61,5% ženy a 38,5% muži. Nemáme k dispozícii údaje o skutočnej štruktúre študentov zúčastnených fakúlt, ale z výsledkov sa ukazuje väčšia ochota odpovedať u respondentov ženského pohlavia. Priemerný vek respondentov bol v SR 20,97 a v ČR 20,83. Percento slobodných je v oboch republikách samozrejme veľké, v SR 98,7% a v ČR 98,1%. Vlastné dieťa malo podľa odpovedí respondentov 3,8% v SR a 1,2% v ČR. Súrodencov nemalo v SR 9% a v ČR 8% odpovedajúcich. 19,3% respondentov prieskumu v SR bolo z Bratislavského kraja a 25,3% v CR z kraja Hlavné mesto Praha.
Dotazník, okrem hore uvedených demografických znakov obsahoval 10 meritórnych otázok zameraných na niektoré súčasné demografické javy.

3. Komparácia názorov mladých v SR a ČR

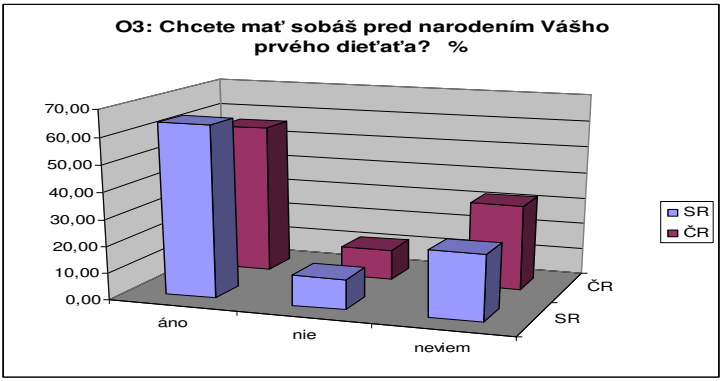
Zaujímavé výsledky prieskumu sme získali na meritórne otázky dotazníka o demografickom správaní.
Na otázku **o1: Koľko detí by ste chceli mať v budúcnosti?** Sme získali odpovede od 0 po 5. v SR bolo 3,8% mladých, ktorí nechcú mať deti a v ČR 3,5%. Priemerný počet detí, ktoré

by chceli mať bol v SR 2,13 a v ČR 2,05. Rozdiel nie je štatisticky signifikantný, čo ukazuje neparametrický Mann-Whitneyov test P=0,201.

Pri otázke **o2: Aký je podľa vás ideálny počet detí v rodine?** Respondenti zo SR odpovedalo v rozmedzí 1 až 4 a v ČR od 0 po 5, pričom počet 0 uviedol jeden respondent, čo predstavuje iba 0,4%. Priemerný ideálny počet detí v rodine podľa respondentov v SR je 2,13 a v ČR 2,18. Čo je štatisticky nesignifikantný rozdiel, P-hodnota Mann-Whitneyovho testu P=0,115.

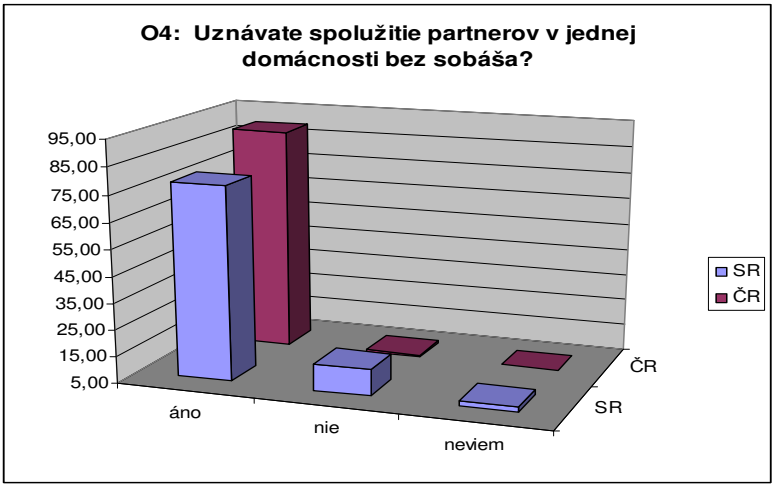
Na otázku **o3: Chcete mať sobáš pred narodením Vášho prvého dieťaťa?** Je rozdelenie odpovedí respondentov v SR a ČR štatisticky nesignifikantne odlišné, P-hodnota Fisherovho exaktného testu P=0,222.

Graf č. 1. Sobáš pred narodením prvého dieťaťa



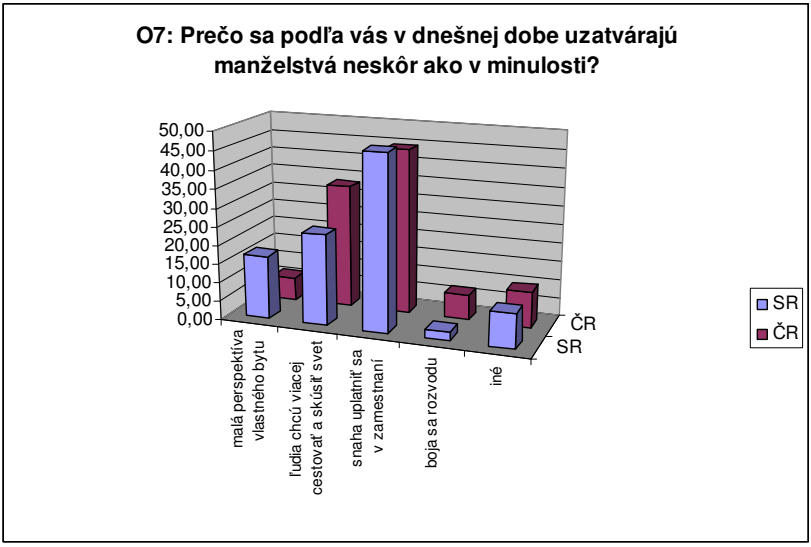
Ďalšia otázka na demografické správanie **o4: Uznávate spolužitie partnerov v jednej domácnosti bez sobáša?** Priniesla nasledujúce odpovede respondentov.

Graf č. 2. Spolužitie partnerov bez sobáša



Pri tejto otázke sú odpovede respondentov signifikantne odlišné, P=0,004. Odpoveď áno je v ČR v 89,3% prípadov, kým v SR v 78,1% a naopak väčšie percento odpovedí nie bolo v SR 15,% ako v ČR 5,7%. Odpoveď neviem bola v SR zastúpená v 6,9% a v ČR v 5% prípadov.
Na otázku **o5 V akom veku by ste chceli uzavrieť manželstvo?** Sme získali odpovede v rozmedzí 21 až 32 rokov s SR a v ČR až po 89 rokov. V ČR odpovedal jeden respondent 89 rokov, taktiež jeden respondent 80 rokov jeden 50 rokov. Tieto odpovede môžu byť recesia, ale na priemernom veku pre uzatvorenie manželstva v ČR 25,11 roka a v SR 27,25 roka majú iba malý vplyv. Rozdiel nie je štatisticky signifikantný.
Otázka **o6: V koľkých rokoch plánujete mať svoje prvé dieťa?** Priniesla odpovede v rozmedzí 20 až 47. Priemerné hodnoty za SR a ČR sú 28,10 a 28,69, čo je štatisticky nesignifikantné, P-hodnota Mann-Whitneyovho testu P=0,202.
Signifikantne odlišné odpovede sme zistili aj pri otázke **o7: Prečo sa podľa vás v dnešnej dobe uzatvárajú manželstvá neskôr ako v minulosti?**, P-hodnota Fisherovho exaktného testu je P=0,002. Prehľadne je štruktúra odpovedí za SR a ČR v grafe č. 3 a podrobne v tabuľke č.1.

Graf č. 3. Príčiny neskoršieho uzatvárania manželstiev v súčasnosti



Signifikantnosť odpovedí spôsobujú najmä odpovede pri variante 1–malá perspektíva vlastného bytu. Respondenti v SR túto alternatívu uvádzali v 16,9% prípadoch a v ČR v menšej miere a síce v 6,1% prípadov. Faktické rozdiely sme zistili aj pri odpovediach 2_ľudia chcú viacej cestovať a skúsiť svet – v ČR 33,2%, kým v SR iba 24,4%, ďalej pri odpovedi 3-snaha uplatniť sa v zamestnaní sme zistili takmer zhodný výsledok – v SR 46,9% a v ČR 44,3%. Odpoveď 4-boja sa rozvodu uviedlo v ČR 6,9% respondentov a v SR iba 2,5%. Zhodný podiel bol zaznamenaný pri odpovedi 5-neviem 9,4% v SR a 9,5% v ČR.

Tabuľka č. 1. Kontingenčná tabuľka komparácie otázky o príčinách neskoršieho uzatvárania sobášov

			o7 Prečo sa podľa vás v dnešnej dobe uzatvárajú manželstvá neskôr ako v minulosti?					Total
			1 malá perspektíva vlastného bytu	2 ľudia chcú viacej cestovať a skúsiť svet	3 snaha uplatniť sa v zamestnaní	4 boja sa rozvodu	5 iné	
stat Štát	1 ČR	Count	16	87	116	18	25	262
		% within stat Štát	6,1%	33,2%	44,3%	6,9%	9,5%	100,0%
		Adjusted Residual	-3,5	1,9	-,5	2,0	,1	
	2 SR	Count	27	39	75	4	15	160
		% within stat Štát	16,9%	24,4%	46,9%	2,5%	9,4%	100,0%
		Adjusted Residual	3,5	-1,9	,5	-2,0	-,1	
Total		Count	43	126	191	22	40	422
		% within stat Štát	10,2%	29,9%	45,3%	5,2%	9,5%	100,0%

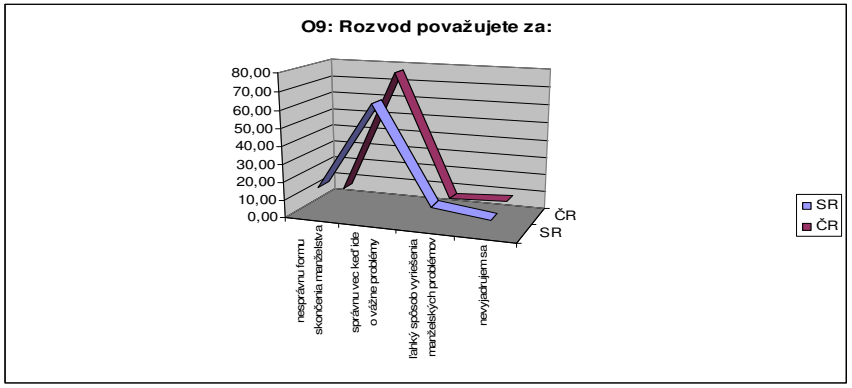
Výsledok za otázku **o8: Z akého dôvodu vstúpite do manželstva vy?** prezentujeme v kontingenčnej tabuľke. Vidno, že percentá odpovedí na jednotlivé varianty sú blízke, čo dokumentuje nesignifikantný rozdiel podľa Fisherovho exaktného testu, P=0,758.

Tabuľka č. 2. Kontingenčná tabuľka komparácie otázky o dôvodoch vstupu do manželstva

			o8 Z akého dôvodu vstúpite do manželstva vy?				Total
			1 lebo chcem mat deti	2 túžim po trvalom manželskom spolužití	3 lebo to tak robia všetci	4 nevstúpim do manželstva	
stat Štát	1 ČR	Count	26	196	10	21	253
		% within stat Štát	10,3%	77,5%	4,0%	8,3%	100,0%
		Adjusted Residual	-,1	-,8	,8	,8	
	2 SR	Count	17	129	4	10	160
		% within stat Štát	10,6%	80,6%	2,5%	6,3%	100,0%
		Adjusted Residual	,1	,8	-,8	-,8	
Total		Count	43	325	14	31	413
		% within stat Štát	10,4%	78,7%	3,4%	7,5%	100,0%

Otázka **o9: Rozvod považujete za:** dáva signifikantne odlišné odpovede respondentov v SR a ČR, P=0,009. Graf č.4 plasticky vyjadruje výsledky, ktoré sú podrobne v tabuľke č. 3.

Graf č. 4. Názory respondentov na rozvod



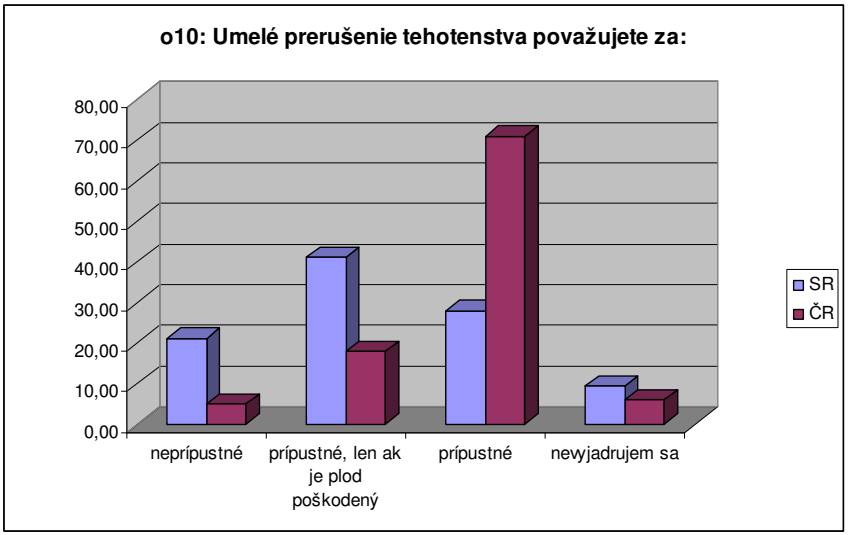
Tabuľka č. 3. Kontingenčná tabuľka komparácie otázky o rozvode

		o9 Rozvod považujete za:				Total
		1 nesprávnu formu skončenia manželstva	2 správnu vec keď ide o vážne problémy	3 ľahký spôsob vyriešenia manželských problémov	4 nevyjadrujem sa	
stat 1 ČR Štát	Count	19	201	19	23	262
	% within stat Štát	7,3%	76,7%	7,3%	8,8%	100,0%
	Adjusted Residual	-2,9	2,7	-1,4	,4	
2 SR	Count	26	103	18	12	159
	% within stat Štát	16,4%	64,8%	11,3%	7,5%	100,0%
	Adjusted Residual	2,9	-2,7	1,4	-,4	
Total	Count	45	304	37	35	421
	% within stat Štát	10,7%	72,2%	8,8%	8,3%	100,0%

Za nesprávnu formu skončenia manželstva považuje rozvod 16,4% respondentov v SR a iba 7,3% v ČR. Ďalší variant odpovede „rozvod je správna vec, keď ide o vážne problémy“ uvádza 64,8% respondentov v SR a 76,7% v ČR. Ľahký spôsob vyriešenia manželských problémov je rozvod u 11,3% respondentov v SR a 7,3% v ČR. Odpoveď nevyjadrujem sa zvolilo 7,5% v SR a 8,8% v ČR.

Posledná otázka **o10: Umelé prerušenie tehotenstva považujete za:** znovu ukázala signifikantne odlišné odpovede za SR a ČR. P- hodnota Fisherovho exaktného testu P=0,000. Výsledky prezentujeme v grafe č. 5 a tabuľke č. 4. Oveľa väčší podiel respondentov za SR považuje umelé prerušenie tehotenstva za neprípustné 21,3%, kým s ČR iba 5,0%. Prípustné len ak je plod poškodený v SR 41,3% a v ČR 17,9. Oveľa nižší podiel v SR je za variant odpovede prípustný 28,1% ako v ČR 71,0%. Nevyjadrilo sa 9,4% v SR a 6,1% v ČR.

Graf č.5. Názory na prerušenie tehotenstva



Tabuľka č. 4. Kontingenčná tabuľka komparácie otázky o umelom prerušení tehotenstva

			o10 Umelé prerušenie tehotenstva považujete za:				Total
			1 neprípustné	2 prípustné, len ak je plod poškodený	3 prípustné	4 nevyjadrujem sa	
stat Štát	1 ČR	Count	13	47	186	16	262
		% within stat Štát	5,0%	17,9%	71,0%	6,1%	100,0%
		Adjusted Residual	-5,2	-5,2	8,6	-1,2	
	2 SR	Count	34	66	45	15	160
		% within stat Štát	21,3%	41,3%	28,1%	9,4%	100,0%
		Adjusted Residual	5,2	5,2	-8,6	1,2	
Total		Count	47	113	231	31	422
		% within stat Štát	11,1%	26,8%	54,7%	7,3%	100,0%

4. Záver

Názory mladých, univerzitných študentov v SR a ČR, na demografický vývoj z aktuálneho prieskumu ukázali zhodu ale aj rozdiely v názoroch respondentov dvoch blízkych republík.

Nesignifikantné rozdiely boli v názoroch na 6 otázok:

o1: Koľko detí by ste chceli mať v budúcnosti?; o2: Aký je podľa vás ideálny počet detí v rodine?; o3: Chcete mať sobáš pred narodením Vášho prvého dieťaťa?; o5 V akom veku by ste chceli uzavrieť manželstvo?; o6: V koľkých rokoch plánujete mať svoje prvé dieťa?; o8: Z akého dôvodu vstúpite do manželstva vy?

Signifikantne rozdielne názory sme zaznamenali pri 4 otázkach: **o4: Uznávate spolužitie partnerov v jednej domácnosti bez sobáša?** V ČR je väčšie pochopenie pre takýto partnerský vzťah**; o7: Prečo sa podľa vás v dnešnej dobe uzatvárajú manželstvá neskôr ako v minulosti?; o9: Rozvod považujete za: a o10: Umelé prerušenie tehotenstva považujete za:** - výsledky poukazujú na konzervatívnejší postoj respondentov v SR.

5. Literatúra

[1]Kanji G. K. (2006): 100 Statistical Tests. 3rd Eddition. SAGE 2006.

[2]Luha, J. (1985): Testovanie štatistických hypotéz pri analýze súborov charakterizovaných kvalitatívnymi znakmi. STV Bratislava 1985.

[3]Luha J. (2009): Matematicko-štatistické aspekty spracovania dotazníkových výskumov.

FORUM STATISTICUM SLOVACUM 3/2009. SŠDS Bratislava 2009. ISSN 1336-7420.

[4]Mládek, J., Širočková, J. (2004). Premeny demografického, najmä správania obyvateľstva na Slovensku. Geografické informácie 8. Stredoeurópsky priestor. Univerzita Konštantína Filozofa, Nitra, pp. 17 – 26.

Adresa autorov:

Ján Luha, RNDr., CSc.
Ústav lekárskej biológie, genetiky a klinickej genetiky LF UK a UNB, Bratislava
jan.luha@fmed.uniba.sk

Dušana Zvarová, 3. ročník
Univerzita Alexandra Dubčeka
Trenčín
dusana.zvarova@gmail.com

Práca bola podporená grantom VEGA 1/0135/09.

Další možnosti modelování úmrtnosti
Another possibilities of mortality modeling

Petr Mazouch, Věra Jeřábková

Abstract: This paper continued on previous papers which focused in basic ways of modeling of mortality. It's known that model based on mortality ratio between individual age groups and development of this indicator in the time period is better way than modeling directly mortality rates, but the models are not so robust. This paper will introduce some ways, how to make the models more robust and useful.

Key words: Mortality, modeling, mortality rate.

Kľúčové slová: Úmrtnost, modelování, míra úmrtnosti

1. Úvod

Tento příspěvek navazuje na předchozí příspěvky (viz [1],[2] a [3]), které se věnovaly možnostem a úskalím modelování úmrtnosti. V předchozích příspěvcích bylo objasněno, proč je vhodné modelovat raději diferenci nebo podíl dvou měr úmrtnosti než přímo míru úmrtnosti. Dochází tak k eliminaci problému, kdy se predikovaná míra úmrtnosti věkové skupiny $x+1$ dostane pod hodnoty úmrtnosti věkové skupiny x.

Také bylo vysvětleno (viz [3]), že výhodnější přístup pro modelování je modelování úmrtnosti v kohortách, než metoda transverzální (tedy průřez všemi kohortami). Základním důvodem je fakt, že se vyhneme potížím s rozdílnou úmrtností mezi kohortami (eliminujeme faktor, který v modelu nelze vystihnout). Model bude tedy založen na longitudinálním přístupu. Tento přístup má však i řadu nevýhod, zejména jeho datová náročnost je značná a v našich zeměpisných šířkách se setkáváme se problémy s datovou základnou v období II. světové války.

Z hlediska volby mezi diferencí a podílem je dle výsledků předchozí studie výhodnější podíl. Při volbě metody diference je často zřetelné, že tato řada je nestacionární. Tato její vlastnost je naprosto logická, protože víme, že míry úmrtnosti rostou s věkem exponenciálně (tato vlastnost je shodná pro všechny kohorty) a tedy spočtené diference mezi měrami úmrtnosti mezi věky x+1 a x v rámci jedné kohorty také ve svých hodnotách rostou s růstem hodnoty x.

Proti tomu hodnoty podílů vykazují alespoň částečnou míru stacionarity, což zajišťuje lepší předpoklady pro další možnosti predikce této řady. Také při hledání trendu vývoje podílu měr úmrtnosti x+1 a x v čase (napříč kohortami) nacházíme lepší výsledky v případě diferencí než v případě absolutních rozdílů.

Jestliže jsme tedy zvolili hodnoty, které budeme dále analyzovat (a které již byly analyzovány v [3]), nezbývá než zjistit, zda výsledky mají očekávané vlastnosti. Při bližším prozkoumání zjistíme, že časové řady podílů (ať již v jedné kohortě pro všechny věkové skupiny nebo pro jednu věkovou skupinu napříč kohortami) vykazují velmi značnou volatilitu, která této řadě, resp. její predikci, jistě nesvědčí.

V následujícím článku se tedy pokusíme jednoduchou metodou tuto volatilitu (zejména extrémy, které řada vykazuje) nějakým způsobem jednoduše odstranit. Jako nejvhodnější způsob se jeví logaritmizace hodnot.

2. Metodika a data

Jednotlivé míry úmrtnosti jsou vypočteny jako podíl počtu zemřelých a středního stavu obyvatelstva příslušné věkové skupiny. Pro věky, kde je již vyšší počet zemřelých (stačí zvolit věkovou skupinu nad 30 let) tedy zřejmě platí

$$m_{x+1,z} \geq m_{x,z}\,, \tag{1}$$

kde x je věk a z je rok narození příslušné kohorty.

Z takto vypočtených měr úmrtnosti můžeme tedy jednoduše spočíst

$$\ln\left(\frac{m_{p,x,z}}{m_{p,x-1,z}}\right) = \ln(r_{p,x,z})\,, \tag{2}$$

kde p je pohlaví, x je věk a z je rok narození příslušné kohorty.

Dle předpokladů z úvodu by takto spočtené logaritmy měly vykazovat jistou míru stability a mělo by tedy nejspíše platit, že

$$\ln(r_{p,x,z}) = \ln(r_{p,x,z+1})\,, \tag{3}$$

kde p je pohlaví, x je věk a z je rok narození příslušné kohorty. Tedy že se změny úmrtnosti vyjádřené jako logaritmus podílu dvou měr nemění v čase (mezi kohortami), což by jako vlastnost umožnilo snadnou predikci měr úmrtnosti do budoucna a také, že platí

$$\ln(r_{p,x,z}) = \ln(r_{p,x+1,z})\,, \tag{4}$$

kde p je pohlaví, x je věk a z je rok narození příslušné kohorty. Tedy, že podíly měr se nemění v rámci kohorty, což je předpoklad známý např. z Gompertz-Makehemovi funkce.

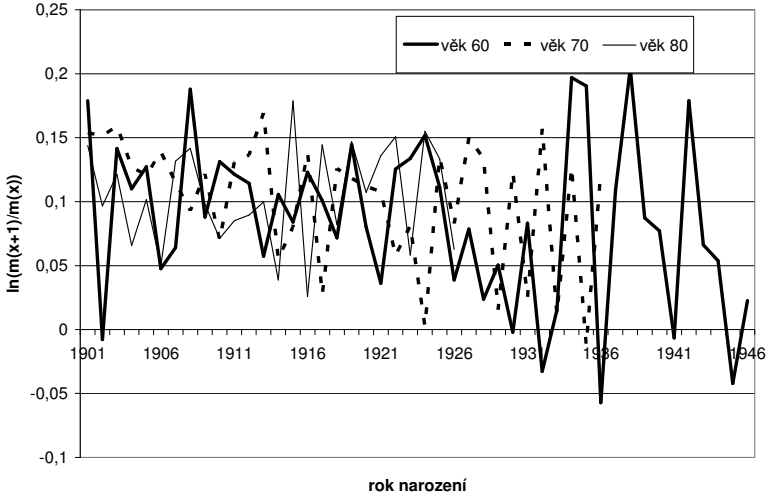
Zdrojová data pro výpočet měr úmrtnosti, tedy počty zemřelých a příslušné střední stavy osob byla převzata z databáze [4]. Data jsou zde pro Českou republiku od roku 1950 (tedy první kompletní kohorta je až od roku 1950, zvolíme-li sledovaný věk vyšší, pak můžeme popivat pochopitelně i kohorty narozené před rokem 1950).

3. Výsledky

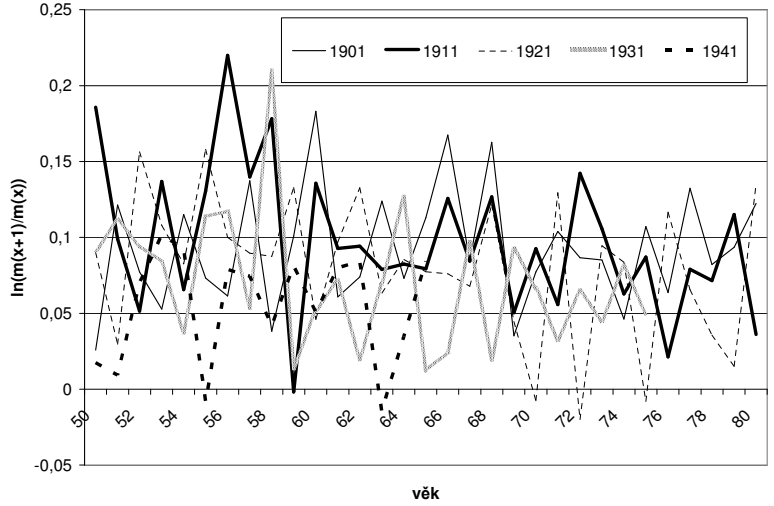
Z prostorových důvodů bylo nutné omezit grafické výstupy pouze na několik ukázek (buď za jednotlivé kohorty nebo za jednotlivé věky). Také datová základna omezuje možnosti výstupu, ale i přes tyto obtíže je patrné, že výsledky po úpravě logaritmem jsou pro další prognózu vhodnější (nelogaritmované hodnoty byly použity ve [2]).

Obrázek 1 ilustruje vývoj logaritmů měr úmrtnosti pro vybrané věky napříč kohortami. Zde by měl platit předpoklad dle vzorce (3), tedy že by tyto logaritmy měly být stabilní napříč všemi kohortami. Je zde patrná velmi vysoká míra volatility. Ta je zcela jistě způsobena výkyvy v úmrtnostech mezi jednotlivými roky (např. chřipkové epidemie), které způsobí prudký nárůst (nebo pokles) úmrtnosti v daném roce, což má zcela jistě vliv na podíl měr úmrtnosti. Celkový vývoj však lze hodnotit na základě tohoto grafického výstupu jako náhodnou oscilaci kolem trendu. Zda je tento trend konstantní a zda je tato oscilace opravdu náhodná musí ověřit matematický postup.

Z obrázku 2 je patrné, že vývoj logaritmů podílů měr úmrtnosti v rámci jedné kohorty je více stabilní. Opět je zde patrná vysoká variabilita, ale na rozdíl od předcházejícího přístupu zde můžeme s naprosto klidným svědomím tvrdit, že jestliže je v jednom roce prudký nárůst míry úmrtnosti, což má za následek velmi nízký logaritmus podílu, pak logicky tato situace musí mít opačný vliv v porovnání s následující mírou úmrtnosti, což odpovídá na otázku, proč se v této řadě střídají propady a vrcholy.



Obrázek 4: Logaritmus podílů měr úmrtnosti vybraných věkových skupin napříč kohortami, Česká republika, ženy



Obrázek 2: Logaritmus podílu měr úmrtnosti vybraných kohort napříč věkem, Česká republika, muži

4. Závěr

Z hlediska dalšího použití je navržený postup vhodný a naprosto logicky musí následovat krok, kde bude nutné matematicky zhodnotit a dokázat vztahy definované zejména vzorci (3) a (4). V případě, že se prokáže konstantní trend těchto hodnot, pak by tato metoda byla zcela jistě vhodná pro využití v případě predikce kohortních měr úmrtnosti, tedy při konstrukci kohortních úmrtnostních tabulek, které se zatím potýkají zejména s nedostatkem datových zdrojů a tedy jakýkoliv jednoduchý systém predikce chybějících hodnot je žádoucí.

V teorii kohortních přístupů k úmrtnosti je základní nevýhodou také to, že je nutné počkat, až celá kohorta vymře. To je také jeden z důvodů, proč by nalezení takové hodnoty „nárůstu" bylo v praxi velmi ceněné a umožnilo by to predikovat vývoje měr úmrtnosti i u kohort, které ještě žijí, což je v současné době, při stále sílící diskusi o důchodové reformě zcela jistě žádoucí (v současnosti se v této problematice využívá transverzálního přístupu, což nevede ke zcela přesným výsledkům).

Závěrem lze tedy poznamenat, že byl učiněn další dílčí krok, který směřuje k cíli vytyčenému na začátku - tedy nalezení souvislostí mezi měrami úmrtnosti v jednotlivých věkových skupinách a jejich vývoji v čase.

Literatúra

[1]MAZOUCH, P. 2009. Modelling of mortality in Czech Republic. Uherské Hradiště 26.08.2009 – 28.08.2009. In: *AMSE 2009 Applications of Mathematics and Statistics in Economy.* Praha : Oeconomica, 2009, s. 289–296. ISBN 978-80-245-1600-4.

[2]MAZOUCH, P. 2010. Modelling of mortality in Czech Republic. In: *AMSE 2010 Applications of Mathematics and Statistics in Economy*. 2010. Banská Bystrica. Sborník v tisku.

[3]MAZOUCH, P. 2010. Možnosti modelování úmrtnosti. *Forum Statisticum Slovacum*, 2009, č. 7, s. 110–114. ISSN 1336-7420

[4]THE HUMAN MORTALITY DATABASE. WEB: WWW.MORTALITY.ORG/

Adresa autora:

Petr Mazouch, Ing., Ph.D.
Vysoká škola ekonomická v Praze
Katedra demografie
Nám. W. Churchilla 4
130 67 Praha 3
mazouchp@vse.cz

Věra Jeřábková, Ing.
Vysoká škola ekonomická v Praze
Katedra ekonomické statistiky
Nám. W. Churchilla 4
130 67 Praha 3
verajerabkova@vse.cz

Příspěvek vznikl za projektu Interní grantové agentury VŠE v Praze „Měření rizik pojišťovny a predikce vstupních parametrů ", projekt č. 29/2010.

Regionálne rozdiely mier nezamestnanosti a miezd na Slovensku a v Českej republike

Regional disparities of the rate of unemployment and wages in Slovakia and Czech Republic

Silvia Megyesiová, Matej Hudak

Abstract: The main objective of regional policy of the European Union (EU) is to achieve a gradual balance of economic, social and other differences between its regions. Comparisons of economic indicators, their development over time and differences of these variables in regions of the EU serve to monitor the needs of separate regions of the EU. In the article we analyse the regional disparities of the rate of unemployment in the regions of Slovakia and Czech Republic on the NUTS3 level. In years 2008 and 2009 we can see a strong influence of the financial crises, the unemployment rate strongly increases in all of the analysed regions. Coefficient of variation was very high during the analysed period of time, it achieved the highest level in Slovakia in 2007.

Key words: regional disparities, unemployment rate, average wages of employee, coefficient of variation.

Kľúčové slová: regionálne rozdiely, miera nezamestnanosti, priemerná mzda zamestnanca, variačný koeficient.

JEL classification: J64, J31, R19

1. Úvod

Hlavným cieľom regionálnej politiky Európskej únie (EÚ) je postupné vyrovnávanie hospodárskych, sociálnych ako aj ďalších rozdielov medzi jej regiónmi. K sledovaniu plnenia tohto cieľa sú krajiny EÚ členené na územia, ktoré zodpovedajú porovnateľnej báze. V roku 2003 vstúpilo do platnosti nariadenie Európskeho parlamentu a Rady č. 1059/2003 zo dňa 26. mája 2003 o vytvorení spoločnej klasifikácie územných štatistických jednotiek NUTS (Nomenclature d'unités territoriales statistiques). Táto jednotná klasifikácia bola zavedená Štatistickým úradom Európskeho spoločenstva Eurostatom v spolupráci s národnými štatistickými inštitútmi každého z 27 členských štátov Únie, v prípade Slovenska v spolupráci so Štatistickým úradom Slovenskej republiky. Klasifikácia NUTS je okrem krajín EÚ zavedená aj v krajinách EFTA (Švajčiarsko, Nórsko, Lichtenštajnsko, Island) a v kandidátskych krajinách (Chorvátsko, Turecko).

Porovnávanie ekonomických ukazovateľov, ich vývoja v čase ako aj rozdiely týchto ukazovateľov v jednotlivých regiónoch krajín EÚ slúžia k monitorovaniu potrieb a vhodnosti podpory konkrétnych regiónov z prostriedkov EÚ.

2. Regionálne rozdiely nezamestnanosti

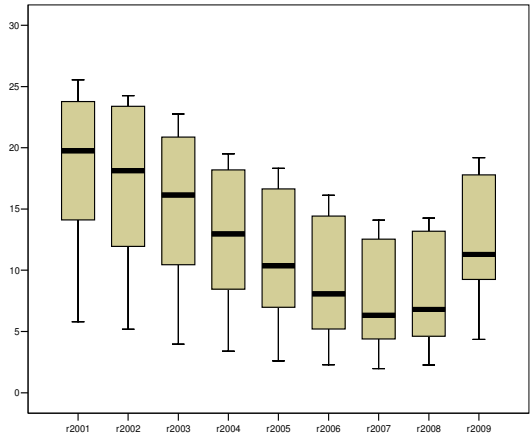
Nezamestnanosť je dôležitým sociálno-ekonomickým faktorom, ktorý výrazne ovplyvňuje sociálnu ako aj ekonomickú úroveň obyvateľstva. *Miera evidovanej nezamestnanosti* sa v súlade s dohovorom Medzinárodnej organizácie práce vypočíta z počtu disponibilných uchádzačov o zamestnanie, ktorí môžu bezprostredne po predložení ponuky vhodného voľného pracovného miesta nastúpiť do zamestnania a z počtu ekonomicky aktívnych osôb za predchádzajúci rok z výberového zisťovania pracovných síl.

Porovnanie mier nezamestnanosti v jednotlivých regiónoch sme uskutočnili na úrovni NUTS3. Zdá sa nám vhodnejšie porovnávať tieto miery na úrovni NUT3 a nie na vyššej úrovni, teda na úrovni NUTS2, vzhľadom k tomu, že týmto porovnaním získame detailnejšie výsledky pre menšie územné celky, ako by tomu bolo v prípade porovnania regiónov iba na úrovni NUTS2.

Slovenská republika je na úrovni NUTS3 rozčlenená na 8 krajov: Bratislavský, Trnavský, Trenčiansky, Nitriansky, Žilinský, Banskobystrický, Prešovský a Košický kraj.

Česká republika je na úrovni NUTS3 rozčlenená na 14 krajov: hlavné mesto Praha, Středočeský, Jihočeský, Plzeňský, Karlovarský, Ústecký, Liberecký, Královéhradecký, Pardubický, Kraj Vysočina, Jihomoravský, Olomoucký, Zlínský, Moravskoslezský kraj. Sledované údaje boli prevzaté z databázy Štatistického úradu SR Regstat ako aj z internetovej stránky Českého štatistického úradu.

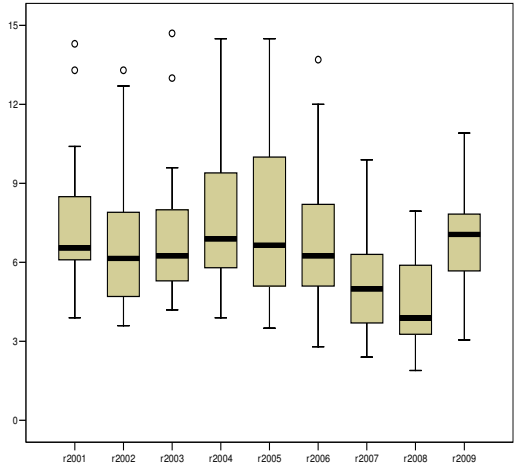
Miera nezamestnanosti v % je prezentovaná vo forme grafov na obrázkoch 1 a 2. V obrázku 1 je zachytený vývoj nezamestnanosti na Slovensku v rokoch 2001 až 2009, pričom z údajov za všetky sledované kraje bol vytvorený box—plot, zachytávajúci tak minimálne, maximálne hodnoty nezamestnanosti, ich dolný a horný kvantil a medián. Z grafu je zrejmé, že postupom rokov dochádzalo k znižovaniu mier nezamestnanosti v jednotlivých krajoch, avšak ich variabilita je vysoká. Kým napríklad v roku 2009 bola miera nezamestnanosti v Bratislavskom kraji na úrovni 4,36 %, v tom istom roku dosiahla v Banskobystrickom kraji 19,19 %. Po obdobiach výrazného poklesu nezamestnanosti je prejav svetovej ekonomickej krízy citeľný aj z pohľadu na tento graf, kde už v roku 2008 a následne v roku 2009 vidieť badateľný rast miery nezamestnanosti v jednotlivých regiónoch Slovenska.



Obrázok 5: Box-plot miery nezamestnanosti v % v SR v rokoch 2001-2009

Vývoj nezamestnanosti v jednotlivých regiónoch Českej republike je zobrazený v obrázku 2. V niektorých obdobiach boli dosiahnuté také úrovne mier nezamestnanosti, ktoré môžeme považovať za extrémne. V roku 2009 bola najnižšia nezamestnanosť v kraji Hlavné mesto Praha s hodnotou 3,1 % a naopak najvyššia miera nezamestnanosti bola dosiahnutá v Karlovarskom kraji, kde vystúpila v sledovanom roku na úroveň 10,9 %. Aj z tohto obrázku 2 je citeľný vplyv svetovej krízy na vývoj mier nezamestnanosti

v jednotlivých regiónoch Českej republiky. Kým v rokoch 2001 až 2008 nezamestnanosť klesala (až na niektoré maximálne hodnoty), v roku 2009 badať citeľný rast miery nezamestnanosti v regiónoch ČR, zvýšili sa tak maximálna, minimálna hodnota miery nezamestnanosti týchto krajov, ako aj ich kvartilové hodnoty.



Obrázok 6: Box-plot miery nezamestnanosti v % v ČR v rokoch 2001-2009

Porovnanie variability mier nezamestnanosti v regiónoch Slovenska a Českej republiky uskutočníme pomocou relatívnej miery variability, ktorá umožňuje jej porovnanie v rôznych súboroch. Variačný koeficient určíme podľa nasledovného vzťahu:

$$V_k = \frac{s}{\bar{x}} \cdot 100 \qquad (1)$$

Hodnoty variačného koeficienta sú tým vyššie, čím je variabilita hodnôt v sledovanom súbore väčšia.

Porovnanie variability pomocou tejto relatívnej miery je zobrazené v tabuľke 1. Regionálne rozdiely sú od roku 2003 silnejšie na Slovensku než v Českej republike. Od roku 2003 sa tieto rozdiely na Slovensku výrazne zvyšujú až do roku 2009. Najvyššiu mieru variability dosiahlo Slovensko v roku 2007. Táto miera teda poukazuje na to, že hlavne v obdobiach, keď Slovensko dosahovalo silný rast ekonomiky, meraný rastom HDP, dochádzalo k silnej regionálnej disproporcii v rámci krajov pri sledovaní miery nezamestnanosti. Naopak v období začiatku krízy v roku 2008 a následne v krízovom roku 2009 dochádza postupne k poklesu variability nezamestnanosti medzi krajmi Slovenska.

V Českej republike bola variabilita nezamestnanosti v jednotlivých krajoch nižšia, okrem rokov 2001 a 2002. Hodnota variačného koeficienta dosahuje v niektorých porovnávaných obdobiach nižšiu hodnotu približne o 10 percentuálnych bodov, znamená to teda výrazne nižšiu variabilitu nezamestnanosti v regiónoch ČR. Túto výrazne nižšiu variabilitu považujeme za pozitívum v porovnaní s rozdielnosťou miery nezamestnanosti v regiónoch Slovenska.

Tabuľka 1: Variačný koeficient mier nezamestnanosti

Rok / Krajina	2001	2002	2003	2004	2005	2006	2007	2008	2009
Slovenská republika	37,8	41,7	44,1	45,6	51,0	56,2	58,7	56,7	42,0
Česká republika	38,3	43,8	41,6	40,9	44,4	43,2	40,4	43,2	32,1

3. Analýza priemernej mzdy zamestnanca

Priemernú hrubú mesačnú mzdu zamestnanca sme podrobili podrobnej analýze za regióny na úrovni NUTS3 tak v Slovenskej republike ako aj v Českej republike a to vo vyjadrení spolu a pre jednotlivé pohlavia. Okrem rozdielov vo výške mzdy z hľadiska regiónov, mzda vykazuje silné rozdiely z hľadiska pohlavia a to vo všetkých sledovaných krajoch. Typické pre obrázok 3 je dosiahnutie extrémnych hodnôt priemernej mzdy obyvateľov Bratislavského kraja vo všetkých sledovaných obdobiach.

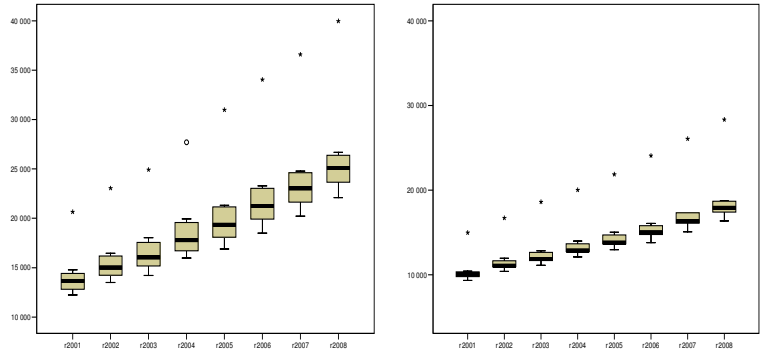
Podiel priemernej mzdy mužov vzhľadom k priemernej mzde žien sa v Bratislavskom kraji pohyboval v období rokov 2001 až 2008 v rozmedzí 38 až 42%. Najnižšie podiely boli dosahované v rokoch 2001-2003, následne sa podiely miezd mužov zvyšovali, vrchol dosiahli v roku 2006.

Najvýraznejší podiel miezd mužov a žien bol dosiahnutý v rokoch 2005 a 2006 v Trnavskom kraji, výška tohto podielu dosiahla hodnoty 46-47%. Po roku 2006 dochádza k postupnému znižovaniu podielu miezd mužov a žien v jednotlivých krajoch SR – viď tabuľku 2.

Tabuľka 2: Podiel priemernej mzdy mužov a žien v percentách

Rok / Krajina	2001	2002	2003	2004	2005	2006	2007	2008
Bratislavský kraj	1.38	1.38	1.34	1.38	1.42	1.41	1.40	1.41
Trnavský kraj	1.37	1.40	1.37	1.44	1.46	1.47	1.41	1.43
Trenčiansky kraj	1.40	1.38	1.38	1.41	1.42	1.43	1.42	1.42
Nitriansky kraj	1.31	1.30	1.29	1.34	1.33	1.36	1.34	1.38
Žilinský kraj	1.36	1.35	1.37	1.37	1.42	1.43	1.42	1.43
Banskobystrický kraj	1.28	1.28	1.26	1.28	1.30	1.32	1.34	1.30
Prešovský kraj	1.31	1.30	1.28	1.32	1.30	1.34	1.34	1.35
Košický kraj	1.42	1.38	1.41	1.43	1.42	1.45	1.43	1.40

Porovnaním variačných koeficientov – obrázok 4 – priemernej mesačnej mzdy regiónov Slovenska a Českej republiky zistíme, že regionálne rozdiely priemernej mzdy krajov Slovenska sú vyššie než regionálne rozdiely miezd jednotlivých regiónov Českej republiky.

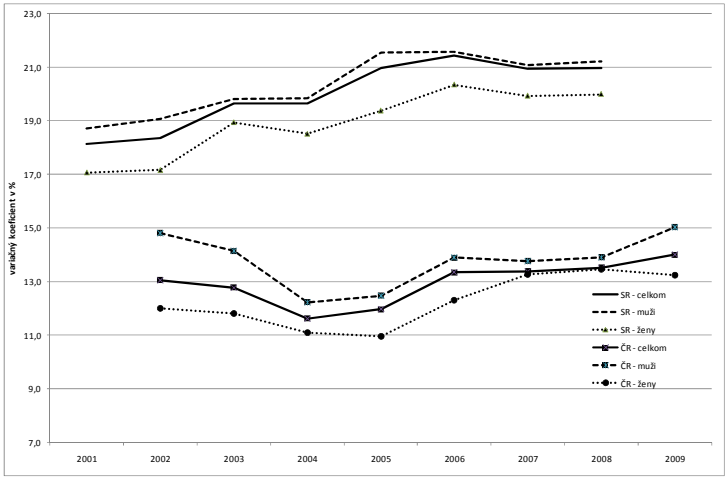


muži *ženy*

Obrázok 3: Hrubá mesačná mzda zamestnanca v SR v Sk

Variačný koeficient miezd slovenských regiónov sa pohybuje približne na úrovni 20%, pričom variačný koeficient je vyšší pri mzdách mužov. Variabilita miezd žien podľa krajov Slovenska dosahuje teda o čosi nižšiu úroveň ako variabilita miezd mužov.

Porovnaním variačného koeficienta v Českej republike vidíme jeho výrazne nižšiu hodnotu, než tomu bolo na Slovensku, variačný koeficient regionálnych miezd ČR dosahuje približne 12-14%. Pri porovnaní pohlaví je situácia podobná ako na Slovensku, vyššia bola variabilita miezd mužov než žien.



Obrázok 4: Variačný koeficient mesačnej mzdy zamestnanca v SR a ČR

4. Záver

Cieľom príspevku bola podrobná analýza regionálnych rozdielov zvolených ukazovateľov na úrovni NUTS3 v Českej republike a na Slovensku. K analýze sme zvolili dva ukazovatele sociálno-ekonomického charakteru a to mieru nezamestnanosti a priemerné mesačné mzdy zamestnancov. Porovnaním týchto ukazovateľov sme zistili, že väčšia variabilita sledovaných ukazovateľov, bola v oboch prípadoch dosiahnutá v regiónoch Slovenska. Regionálne rozdiely sa v čase hospodárskej krízy čiastočne znížili a to na Slovensku ako aj v Českej republike.

5. Literatúra

[1]MEGYESIOVÁ Silvia. Štrukturálne ukazovatele hodnotenia Lisabonskej stratégie. In *Forum Statisticum Slovacum*. ISSN 1336-7420. Roč. 2, č.4 (2006), s. 139-145.

[2]MEGYESIOVÁ, Silvia. Nezamestnanosť na Slovensku a v okolitých krajinách. In *Acta oeconomica Cassoviensia No 3*. Košice : Podnikovohospodárksa fakulta EU so sídlom v Košiciach, 1999. ISBN 80-88964-15-6. s. 303-308. [vedecký redaktor: Viktória Bobáková].

[3]EGYESIOVÁ, Silvia. Viackriteriálna analýza ekonomík krajín Európskej únie. In *Sborník příspěvků, Mezinárodní statisticko-ekonomické dny na VŠE v Praze*, *3. ročník vědecké konference*. ISBN 978-80-86175-66-9. Zborník na CD.

[4]Regional Disparities and Cohesion: What Stategies for the Future, IP/B/REGI/IC/2006_201, prevzaté z: http://www.europarl.europa.eu/hearings/20070625/regi/study_en.pdf

[5]Stankovičová, I. : Štatistická analýza krajín sveta podľa vybraných demografických ukazovateľov v roku 1998. In: Demografické, zdravotné a sociálno-ekonomické aspekty úmrtnosti. - Bratislava: SŠDS, 1999. - ISBN 80-88946-00-X.

[6]Stankovičová, I.: Analýza rozptylu a systém SAS. In: Forum Statisticum Slovacum. - Roč. 1, č. 2 (2005), s. 125-129

[7]Vojtková, M.: Hodnotenie zamestnanosti v regiónoch krajín V 4. In: Ekonomické rozhľady. 2/2010, EU v Bratislave, s. 192-206. ISSN 0323-262X.

[8]Databáza ŠÚ SR Regstat: http://px-web.statistics.sk/PXWebSlovak/

[9]Český štatistický úrad: http://www.czso.cz/csu/redakce.nsf/i/regiony_mesta_obce_souhrn

Adresa autorov:

Silvia Megyesiová, Ing., PhD.
Podnikovohospodárksa fakulta EU
Tajovského 13
041 30 Košice
megyesiova@euke.sk

Matej Hudák, Ing., PhD.
Podnikovohospodárksa fakulta EU
Tajovského 13
041 30 Košice
matej.hudak@euke.sk

Príspevok bol spracovaný v rámci projektu VEGA č. 1/0339/10.

Inovácie v zbere a spracovaní štatistických údajov prostredníctvom RFID
Innovations in collecting and processing statistical information through RFID

Branislav Mišota, Adam Sorokáč

Abstract: In our article we will try to bring the model example of the draft structure and bind themselves to retrieve data into the SAP system RFID technology.

Key words: radio-frequency identification (RFID), client server, tag, flow of goods

Kľúčové slová: rádio-frekvenčná identifikácia (RFID), značka, klient server, tok výrobkov

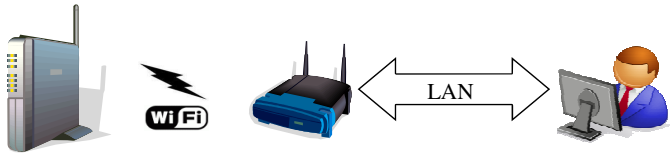
JEL classification: M11, M15, L62

1. Úvod

V tomto článku naviažeme na náš príspevok publikovaný vo Forum Statisticum Slovacum, č. 4 / 2010 (FSS 4/2010), ktorý popisoval súčasný stav štatistickej identifikácie vo výrobnom procese producenta subkomponentov, pre brzdové systémy automobilov a boli v ňom navrhnuté aj možnosti, prostredníctvom ktorých je možné zvýšiť kvalitu získaných štatistických údajov v podnikových manažérskych informačných systémoch, prostredníctvom zavedenie novej tzv. RFID technológie v niekoľkých samostatne funkčných blokoch spojených pomocou bezdrôtovej technológie WiFi do jedného hierarchického systému. Najvyššie postaveným prvkom voči jednotlivým blokom bude server, ktorý bude prijímať a dodatočne spracovávať prijaté údaje. Prvotné predspracovanie bude vykonané v blokoch uvedených vo FSS 4/2010. Server bude prijaté údaje ďalej triediť a v konečnej fáze usporadúvať podľa zvoleného predpisu do tabuliek, s ktorými bude ďalej pracovať logistika alebo iný správca výroby.

2. Spracovanie údajov

Blok spracovania údajov bude zastupený serverom, ktorý je zároveň i hierarchicky nadradeným voči dvom vyššie popísaným blokom. Pod pojmom server rozumieme počítačovú zostavu bez výstupných periférií určenú len na spracovávanie údajov a komunikáciu. Zároveň bude slúžiť ako rozhranie medzi najnižšou vrstvou a horizontálne nadradenou (viď Obr. 1). Všetky konečné úkony vykonané buď na jednotlivých pracoviskách cez operačné panely (ďalej OP - priemyselné počítače s dotykovým displejom) alebo na čítačkách pri vstupe do haly budú odoslané serveru. Ten, ako nadradený správca systému, bude následne spracovávať prijaté informácie (zatrieďovať podľa dohodnutého predpisu). Zároveň po identifikácii pracovníka bude zdieľať jeho údaje s OP (viď Obr. 2), taktiež jeho úlohou bude časová synchronizácia s jednotlivými pracoviskami (OP).



Obr.1 Komunikačná schéma bloku spracovania údajov.

3. Predpis zápisu údajov

Pre maximálny výkon celého systému je práve predpis najdôležitejší. Od neho bude závisieť celková prehľadnosť údajov, čiže čitateľnosť údajov logistikou. Daná firma používa na správu a prehľad výroby komplexný softvérový produkt SAP. Práve táto skutočnosť už hovorí o predpise ukladania dát. Preto bude potrebné ďalej zabezpečiť i samotné načítanie dát do vyššie spomenutého informačného systému. Zároveň bude na serveri vytvorená databáza prijatých údajov z dôvodu spätnej kontroly v prípade reklamácie zo strany odberateľa.

Potrebný obsah údajov na značke:

- čas a dátum kedy bola operácia vykonaná (začiatok a koniec)
- číslo zmeny v rámci dňa (pracuje sa na tri 8-hodinové zmeny, kde 1- ranná, 2- poobedná a 3-nočná)
- meno pracovníka, ktorý vykonal pracovný úkon
- druh pracovnej operácie, ktorý bol vykonaný na výrobku (frézovanie, opracovanie, pokovovanie)
- názov stroja, na ktorom bola prevedená prac. operácia
- počet kusov, ktoré boli opracované (počty sú zvyčajne dané vopred a to predpisom na ukladanie výrobkov)
- počet kusov, ktoré sa poškodili alebo už boli poškodené

4. Forma zápisu údajov

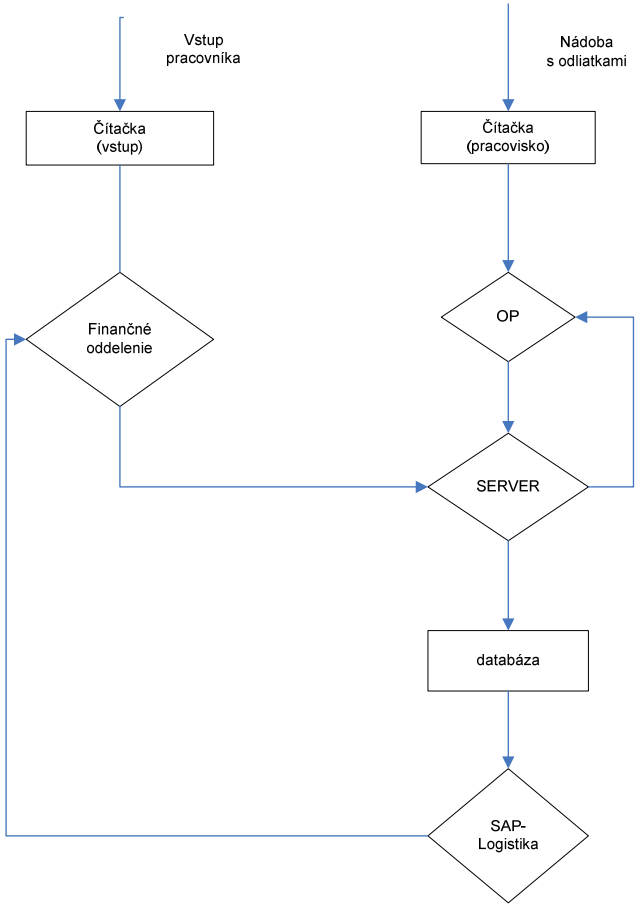
Výstupné informácie zo značiek budú vo forme číselných kódov zaužívaných uvedenou firmou, taktiež z dôvodu korektného čítania informačným systémom SAP. Samozrejme v databáze budú tieto číselné kódy preložené priamo do textu pre lepšiu čitateľnosť údajov.

Tab.1 Návrh zápisu údajov do databázy (server)

dátum	zmena	začiatok operácie	koniec operácie	priezvisko meno	druh operácie	názov stroja	typ výrobku	ks	pošk. ks
11.4.2009	1	9:10:56	9:40:50	Tkac.Jan	Frezovanie	Honsberg	doštičky	90	0
	1	9:42:58	10:05:34	Tkac.Jan	Frezovanie	Honsberg	doštičky	90	0
	1	10:08:24	10:40:45	Tkac.Jan	Frezovanie	Honsberg	doštičky	90	0
	1	10:43:43	11:10:30	Tkac.Jan	Frezovanie	Honsberg	doštičky	90	0

údaje zadávané ručne pracovníkom
OP
Server

Väčšina výstupných údajov v databáze bude získaná z OP, kam môžeme zarátať aj skupinu dát zadávaných ručne pracovníkom. Jediná informácia získaná mimo OP je meno pracovníka, pričom v niektorých prípadoch to môže byť vylúčené, ak nastane prechod pracovníka na iné pracovisko, kedy sa prihlasuje priamo na OP. Jednotlivé údaje v určenom predpise, ako už bolo spomenuté, budú zasielané hlavne správcovi, v našom prípade rozumieme logistiku podniku, čiže systém SAP, a taktiež pre nutnosť spätnej kontroly sa budú údaje archivovať na serveri v navrhnutom predpise (viď Tab. 1).



Obr.2 Príjem a spracovanie údajov serverom

5. Záver

V našom článku sme snažili na vzorovom príklade priblížiť návrh štruktúry a naviazanie na samotné načítanie dát do informačného systému SAP v prípade zavedenia technológie rádiofrekvenčnej identifikácie komponentov ako náhrady klasických čiarových kódov. Prostredníctvom takéhoto inovatívneho prístupu je možné nielen zvýšiť kvalitu získaných štatistických údajov v podnikových informačných systémoch ale aj efektívnejšie vytvárať databázu prijatých údajov z dôvodu spätnej kontroly v prípade reklamácie zo strany odberateľa.

Literatúra

[1] Chajdiak. J. (2009). Štatistika v exceli 2007. Bratislava Statis 2009, ISBN 978-80-85659-49-8.

[2] HOLOMEK, J., ŠIMANOVSKÁ T. 2002. Úvod do metodológie praktických vied. Bratislava 2002.163 s. ISBN 80-8054-249-X.

[3] MICHÁLEK, Ivan - VACULÍK, Juraj - KOLAROVSZKI, Peter. Integrácia *RFID* do oblasti logistiky. In Ekonomicko-manažérske spektrum : vedecký časopis Fakulty prevádzky a ekonomiky dopravy spojov Žilinskej univerzity v Žiline. - Žilina : Fakulta prevádzky a ekonomiky dopravy a spojov Žilinskej univerzity, 2009. ISSN 1337-0839, 2009, roč. 3, č. 1, s. 44-55.

[4] GABRIŠ, Miloš. Nasadenie *RFID* pri zvyšovaní efektívnosti distribučných procesov. In Merkúr 2008 : výsledky vedeckej práce mladých vedeckých pracovníkov [elektronický zdroj]. - Bratislava : [Obchodná fakulta EU], 2008. ISBN 978-80-225-2770-5, s. 63-67. VEGA 1/4598/07.

[5] Od čiarového kódu k *RFID* a ďalej. In Doprava a logistika : odborný mesačník vydavateľstva Ecopress. - Bratislava : ECOPRESS, 2007. ISSN 1337-0138, November 2007, roč. 2, č. 11, s. 46-47.

Adresa autorov:

Branislav Mišota, Ing
ÚM STU – OEMP
Vazovova 5
812 43 Bratislava
branislav.misota@stuba.sk

Adam Sorokáč, Bc.
FEI STU
Ilkovičova 3
812 19 Bratislava
asorokac@gmail.com

Príspevok bol spracovaný v rámci riešenia úlohy VEGA č. 1/0536/10 Inovácie ako strategický základ konkurenčnej schopnosti SR (Smerovanie, meranie a podpora inovačných procesov).

Zhlukovanie ako štatistická analýza v ekonomických vedách
Statistical analysis through cluster analysis in economic sciences

Ladislav Mura

Abstract: In this paper we deal with one of the possibilities provided by statistical analysis, i.e. clustering. In the paper the problems are identified and the advantages of the application of cluster analysis and the possibilities of its usage in economic sciences are described.

Key words: Statistical analysis, cluster analysis, cluster, data

Kľúčové slová: štatistická analýza, zhlukovanie, zhluk, dáta

JEL classification: C10, C38

1. Úvod

Zhluková analýza sa zaoberá tým, ako by mali byť objekty, teda štatistické jednotky zaradené do skupín tak, aby bola čo najväčšia podobnosť v rámci skupín a čo najväčšia rozdielnosť medzi skupinami. Zhluková analýza sa používa v rôznych oblastiach ekonomických vied. Premennými kritériami môžu byť pohlavie, vek, vzdelanie, životný štýl, náboženstvo, skúsenosti s produktom, veľkosť spotreby, frekvencia spotreby a podobne. [5]

Štatistická analýza formou zhlukovania je súborom metód, ktoré umožňujú hľadať v empirických údajoch zoskupenia podobných objektov. Použitie metód zhlukovej analýzy je obzvlášť vhodné najmä v štádiu explorácie problému. Na rozdiel od faktorovej analýzy sa pri uplatnení zhlukovej analýzy zväčša nevenuje pozornosť základným dimenziám popisu sledovaných javov (premenným), ale základným typom sledovaných javov ako takých (objektom), ich podobnosti a nepodobnosti (hoci je niekedy problematické vymedziť, čo vlastne podobnosť je). Zhluková analýza sa teda využíva na hľadanie typov objektov, ktoré sa vyznačujú špecifickými charakteristikami, tzn. slúži ako primeraný nástroj generovania predpokladov o klasifikácii objektov.

Pri analýze a komparácii rôznych ukazovateľov úspešnosti podnikateľských subjektov sa okrem iných štatistických metód často využíva práve zhluková analýza. Pomocou tejto metódy je možné podniky začleniť do zhlukov tak, aby v rámci skupiny podnikateľských subjektov bola čo najväčšia podobnosť. Na uskutočnenie tejto analýzy je možné využiť špecifický štatistický software. [1]

2. Cieľ, materiál a metódy

Cieľom článku je poukázať na štatistickú analýzu prostredníctvom zhlukovania a možnosti jej využitia vo výskume v ekonomických vedách. Parciálnym cieľom je objasniť základné vlastnosti zhlukovej analýzy a ukázať praktickú aplikáciu zhlukovej analýzy s využitím počítačového programu. Zo softvérových produktov sme využili programy Mathematica a Matlab. Pri spracovaní príspevku sme využili sekundárne literárne pramene. Použité boli predovšetkým logické metódy.

3. Teoretické východiská

Základným cieľom zhlukovej analýzy je redukcia, teda zjednodušenie údajov. Ak sa totiž v údajoch nájde určitá štruktúra typov, môžu byť tieto jednoduchšie popísané na základe spoločných charakteristík svojich členov a môžu byť pomenované. Zhluková analýza je tak vhodným prostriedkom k vytvoreniu určitej typológie, redukcie veľkého množstva objektov

do tried, a to podľa tých hľadísk, ktoré sú pre daný problém relevantné. Jednotlivé triedy, resp. objekty do nich prináležiace sú zväčša charakterizované multidimenzionálne, tzn. objekty sú triedené na základe viacerých charakteristík (premenných).

Pri zohľadnení iba jednej premennej (1-D) je nájdenie zhlukov veľmi jednoduché, keďže sa hodnoty premennej nanesú na číselnú os a zhluky sa identifikujú vizuálne (napr. podľa veku nájdeme v súbore 2 skupiny respondentov: jednu okolo 15 rokov a druhú okolo 40 rokov). Podobne použitím X-Y grafu možno jednoducho identifikovať zhluky pri zohľadnení 2 premenných (2-D). V priestore (3-D) sa pomocou interaktívneho X-Y-Z grafu tiež dajú nájsť zhluky vizuálne. Vizuálne identifikovať zhluky pri zohľadnení viac ako 3 premenných súčasne sa však už nedá. Práve vtedy sa používa zhluková analýza. [2]

Podstatnou charakteristikou zhlukovej analýzy je skutočnosť, že triedy, resp. typy objektov (zhluky) nie sú a priori známe (práve tým sa zhluková analýza líši od iných metód klasifikácie). Zhluky sú v priebehu samotnej analýzy odvodzované z údajov, a to tak čo sa týka ich počtu, ako aj čo sa týka ich charakteristík. Z predchádzajúceho vyplýva, že ide o vysoko empirickú metódu, a teda že rôzne postupy uplatnené na rôznych údajoch môžu viesť aj k rôznym typológiám, t. j. k rôznej interpretácii dát.

Metódy zhlukovej analýzy teda smerujú k vytvoreniu zhlukov – tried objektov tak, aby vzájomná podobnosť objektov zaradených do jedného zhluku bola čo najväčšia, a zároveň aby sa maximalizovala medzizhluková nepodobnosť. K uchopeniu podobnosti, resp. nepodobnosti sa používajú rôzne miery – koeficienty podobnosti. Vo všeobecnosti rozlišujeme miery podobnosti, ktoré nadobúdajú tým väčšie hodnoty, čím sú si sledované objekty podobnejšie (rôzne druhy korelácií) a miery nepodobnosti, ktoré svojou rastúcou hodnotou signalizujú zmenšujúcu sa podobnosť medzi objektmi (rôzne druhy dištancií, resp. vzdialeností).

Zhluková analýza zahŕňa množstvo metód. Rozlišujeme dve základné skupiny [4]:

- Hierarchické zhlukovacie metódy,
- Nehierarchické zhlukovacie metódy

Hierarchické zhlukovacie metódy vychádzajú z jednotlivých objektov, ktoré reprezentujú zhluky, pričom každý jeden objekt tvorí prvotný zhluk. Ich spájaním sa v každom kroku počet zhlukov postupne zmenšuje až sa nakoniec všetky zhluky spoja do jedného celku. Hierarchické metódy vedú k hierarchickej (stromovej) štruktúre, ktorá sa graficky zobrazuje ako stromový diagram (dendrogram). Stromové zhlukovacie metódy začínajú výpočtom vzdialenosti medzi objektmi, ktoré nazývame ako Euklidovská vzdialenosť. Alternatívnu vzdialenosť prestavuje vzdialenosť Manhattan (City-block). Euklidovská vzdialenosť vyjadruje vzdušnú vzdialenosť medzi dvoma objektmi a vzdialenosť Manhattan najkratšiu vzdialenosť, ktorú musí napr. chodec prejsť aby sa v meste dostal z jedného miesta na druhé. Výhoda vzdialenosti Manhattan spočíva v znížení dopadu extrémnych prípadov (outliers) na výsledky. Je treba uviesť, že existujú ešte aj iné typy vzdialeností, ktoré sa používajú napr. pri kategorických premenných. Po vypočítaní vzdialenosti medzi všetkými dvojicami objektov musíme určiť pravidlo podľa ktorého sa budú objekty spájať do zhlukov, teda ako sa bude určovať vzdialenosť medzi zhlukmi.

Existujú viaceré pravidlá spájania, z ktorých vyberáme nasledovné:

- *Single linkage* (Nearest Neighbour) – jednoduché spájanie (najbližší sused). Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť dvoch najbližších členov.
- *Complete linkage* (Furthest Neighbour) – kompletné spájanie (najvzdialenejší sused). Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť dvoch najvzdialenejších členov.

- *Unweighted pair-group average* (Group Average) – nevážený párový priemer (priemer skupín). Vzdialenosť medzi zhlukmi je definovaná ako priemerná vzdialenosť medzi všetkými pármi, pričom 1.člen je z 1.zhluku a 2.člen z 2.zhluku.
- *Weighted pair-group average* (Simple Average) – vážený párový priemer (jednoduchý priemer). Podobná ako predošlá z tým rozdielom, že veľkosti zhlukov (počty objektov) sa berú ako váhy.
- *Unweighted pair-group centroid* (Centroid) – nevážený centroid (centroid). Vzdialenosť medzi dvoma zhlukmi je definovaná ako vzdialenosť centroidov týchto dvoch zhlukov. Centroid je vektor priemerov (každá súradnica je priemer príslušných súradníc objektov v zhluku).
- *Weighted pair-group centroid* (Median) – vážený centroid (medián). Podobná ako predošlá z tým rozdielom, že veľkosti zhlukov (počty objektov) sa berú ako váhy.
- *Wardova metóda.* Táto metóda sa zreteľne odlišuje od všetkých ostatných pretože na určenie vzdialenosti medzi zhlukmi využíva prístup analýzy rozptylu. S touto metódou sa zhluky vytvárajú tak, aby sa vnútrozhlukový súčet štvorcov minimalizoval.

Nehierarchické zhlukovacie metódy nevytvárajú stromovú štruktúru. Najznámejšia nehierarchická zhlukovacia metóda je metóda k-priemerov (k-means). Táto metóda sa vyznačuje tým, že vyprodukuje presne k-zhlukov tak, aby bol vnútroskupinový súčet štvorcov minimálny. Najvhodnejšia je na formovanie malého počtu zhlukov z veľkého počtu pozorovaní. Vyžaduje však intervalové premenné bez extrémnych hodnôt (outliers). Nominálne premenné sa dajú použiť ale môžu spôsobovať problémy. Užitočnou metódou je neurčité zhlukovanie (Fuzzy clustering) [3], ktoré na rozdiel od ostatných zhlukovacích metód, umožňuje čiastočné zaradenie objektu do viacerých zhlukov a to pomocou pravdepodobnosti. Cieľom je zabrániť skresleniu zhlukovania kvôli prítomnosti nezaraditeľných objektov. Takéto indivíduum sa nepriradí ku žiadnemu zhluku (od každého sa príliš odlišuje), ale priradia sa mu pravdepodobnosti s ktorými sa bude nachádzať v jednotlivých zhlukoch. Metóda sa často používa pri odhaľovaní podvodov v rôznych oblastiach. Napr. v bankovníctve sa bez vopred formulovanej definície podozrivej operácie z miliónov operácií klientov identifikuje pár desiatok takých, ktoré sa od zvyšných (zoskupených do niekoľkých zhlukov) pri použití viacerých premenných (napr. obrat, typ operácie, konštantný symbol, čas od zadania po jej splatnosť atď.) výrazne odlišujú.

Podnetnou sa javí možnosť validizácie typológie získanej prostredníctvom nehierarchickej zhlukovej analýzy, ktorá má východisko v znovuprevedení zhlukovej analýzy ešte na dvoch častiach výskumného súboru, pričom rozdelenie na dané dve časti je náhodné. Za účelom validizácie výsledku nehierarchickej zhlukovej analýzy na celom súbore sa odporúča prostredníctvom hierarchickej zhlukovej analýzy porovnať tento celosúborový výsledok s čiastkovými výsledkami nehierarchických zhlukových analýz prevedených na dvoch náhodných častiach súboru. Na tri takto získané typológie (profily typov) sa aplikuje hierarchická zhluková analýza druhého rádu. Pomocou nej sa takto vlastne určí podobnosť medzi typmi dvoch čiastkových typológii a celosúborovej typológie. Ak sú vo výslednom dendrograme na „najbližšej úrovni“ spojené vždy trojice príslušných typov (z celého súboru, z prvej časti súboru a z druhej časti súboru), môžeme výslednú celosúborovú typológiu považovať za validnú.

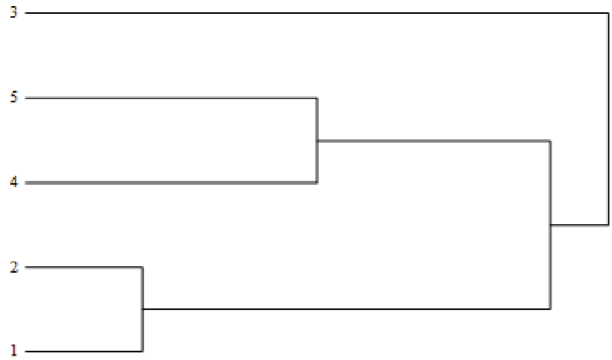
4. Praktické využitie zhlukovej analýzy s počítačovou podporou

V nasledujúcej časti článku predstavíme dva počítačové programy (Mathematica a Matlab), v ktorých je možné uskutočniť zhlukovú analýzu.

Program Mathematica ponúka pre zhlukovú analýzu funkcie Hierarchical Clustering Package. Prvým príkladom, ktorý si ukážeme bude výpočet matice vzdialenosti. Pred výpočtom je nutné zaviesť balíček HierarchicalClustering, príkazom Needs["HierarchicalClustering`"]. Príkaz pre výpočet matice vzdialeností je Distance-Matrix[data]. Príkaz vypočíta maticu vzdialeností cez použitie Euklidovej vzdialenosti. Ak by sme chceli použiť inú vzdialenosť, napríklad manhattan, musíme upraviť príkaz na tvar DistanceMatrix[data,Distance-Function -> ManhattanDistance]. Výraz MatrixForm zabezpečí, aby bol výsledok v tvare matice.

Ďalším príkladom je zostrojenie dendrogramu. K tomuto používa program Mathematica funkciu DendrogramPlot[data]. Zadáme príkaz DendrogramPlot [data,DistanceFunction->SquaredEuclideanDistance, Orientation->Right,Linkage->"Single", LeafLabels->Automatic], kde DistanceFunction -> SquaredEuclideanDistance je nami zvolená vzdialenosť. Orientation -> Right zabezpečí, aby os body bola vľavo. Link-age -> "Single" určuje, ktorá metóda spojenia bude použitá. LeafLabels ->Automatic slúži na to, aby se u každého listu stromu zobrazil popis. V našom prípade je to poradové číslo v dátovej matici data. Náš dendogram znázorňuje obrázok 1.

Obrázok 1 Dendogram v programe Mathematica

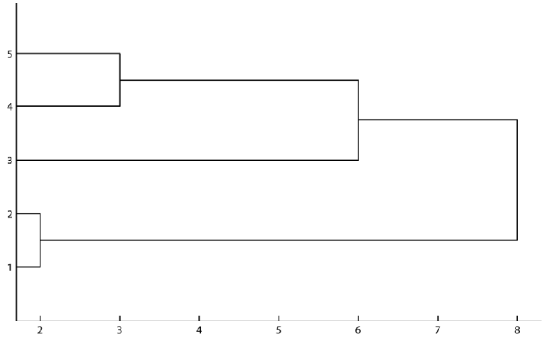


Zdroj: výstup zo softvéru Mathematica

Program Matlab má zabudovanú zhlukovú analýzu v Statistics Toolbox. Z tohto balíku využijeme časť Cluster Analysis. Dáta budú rovnaké, aby sme následne mohli programy porovnať. Prvým príkladom bude teda výpočet matice vzdialenosti. Matlab má pre jej výpočet funkciu pdist(data), ako vzdialenosť je štandardne nastavená Euklidova vzdialenosť. Ak chceme použiť maticu vzdialenosti našich údajov pri použití vzdialenosti manhattan, bude príkaz v tvare squareform(pdist(data,'cityblock')). Squareform v príkaze zaistí, aby vzdialenosti boli zobrazené maticovo.

Druhým príkladom je zostrojenie dendrogramu. Matlab pre túto možnosť funkciu dendrogram(z). Vzdialenosť bude počítaná za použitia vzdialenosti manhattan. Za metódu spojenia bude zvolená metóda najvzdialenejšieho suseda. Príkazy na zostrojenie dendrogramu sú nasledovné: mv = pdist(data,'cityblock');z = linkage(mv,'complete'); dendrogram(z,'orientation','right');. Prvý príkaz spočíta maticu vzdialeností. Druhý príkaz vytvorí hierarchický strom zhlukov z matice vzdialeností. V poradí tretí príkaz slúži na zostrojenie dendrogramu. Vďaka príkazu v tvare 'orientation','right' bude mať dendrogram začiato vľavo. Výstup ilustruje obrázok 2.

Obrázok 2 Dendogram v programe Matlab



Zdroj: výstup zo softvéru Matlab

Komparácia počítačových programov pre zhlukovú analýzu:

Skúmali sme, ktorý z uvedených softvérových vybavení počítača je pre zhlukovú analýzu lepší. Prvé porovnanie za zamerané na vzdialenosti. Tabuľka 1 znázorňuje oba programy a vzdialenosti, ktoré umožňujú.

Tabuľka 1 Porovnanie počítačových programov pre vzdialenosti

Mathematica	Euclidean	SquaredEuclidean	Manhattan	Chebyshev
Matlab	Euclidean	SEuclidean	CityBlock	Chebyshev
Mathematica	BrayCurtis	Canberra	Cosine	Correlation
Matlab	Mahalanobis	Minkowski	Cosine	Correlation

Zdroj: softvér Mathematica a Matlab

Z tabuľky 1 je vidieť, že obidva produkty ponúkajú dostatok vzdialeností. Pri SEuclidean v programe Matlab nie je štvorec Euklidovej vzdialenosti, avšak štandardizovaná Euklidova vzdialenosť. V tabuľke 1 sú uvedené len miery pre meranie vzdialenosti medzi číselnými hodnotami. Oba počítačové programy ponúkajú aj miery vzdialeností (nepodobnosti) pre logické hodnoty a tiež pre reťazce. Napr. Hammingovu vzdialenosť, Jaccardovu nepodobnosť, Russell-Raova-u nepodobnosť a ďalšie. V oboch programoch je rovnako možné nadefinovať si aj vlastnú mieru vzdialenosti. K meraniu vzdialenosti patrí i určenie metódy spojenia. Tieto metódy zaznamenáva tabuľka 2.

Tabuľka 2 Porovnanie počítačových programov pre metódy spojenia

Mathematica	Single	Average	Complete	Weighted	Centroid	Median	Ward
Matlab	Average	Centroid	Complete	Median	Single	Ward	Weighted

Zdroj: softvér Mathematica a Matlab

Z tabuľky 2 je vidieť, že obidva produkty ponúkajú užívateľovi rovnaké spojovacie metódy. Softvér Mathematica umožňuje tiež nadefinovať si vlastnú spojovaciu metódu. Význam jednotlivých metód je nasledovný: Single - metóda najbližšieho suseda, Average – metóda priemernej väzby, Complete - metóda najvzdialenejšieho suseda,Weighted – vážená metóda priemernej väzby, Centroid - metóda centroidov, Median – metóda mediánu, Ward - Wardova metóda.

4. Záver

V predkladanom článku sme venovali pozornosť objasneniu štatistickej analýzy prostredníctvom zhlukovania, vymedzili sme rozdelenie zhlukovej analýzy, definovali sme možnosti, ktoré nám poskytuje. Poukázali sme na niektoré praktické sféry jej využitia v ekonomickom výskume s podporou vybraných počítačových programov Mathematica a Matlab. Záverom môžeme konštatovať, že prezentované softvérové produkty sú vhodné pre uskutočnenie zhlukovej analýzy.

5. Literatúra

[1] MURA, L. 2010. Skúmanie internacionalizácie podnikania v mliekarenskom priemysle Slovenska. Praha: Vysoká škola ekonomická, 2010, zborník v tlači.

[2] STEHLÍKOVÁ, B.2003. Štatistická analýza systémom SAS. Nitra: SPU, 2003, 127s. ISBN 80-8069-221-1

[3] STEHLÍKOVÁ, B. 2004. Fuzzy zhluková analýza ako prostriedok hodnotenia rozdielov. In: Zborník príspevkov z konferencie „12. Slovenská štatistická konferencia - Štatistika a integracia“. Bratislava : SŠDS, 2004, s. 165-169. ISBN: 80-88946-37-9

[4] STEHLÍKOVÁ, B. Zhluková analýza. Metodická príručka. Nitra: SPU, 1999

[5] RIMARČÍK, M. [online] [cit. 2010-10-18] Dostupné na internete: http://rimarcik.com/navigator/ca.html

Adresa autora:

Ladislav Mura, Ing. et Bc., PhD.
Ústav odborných predmetov a IT
Dubnický technologický inštitút v Dubnici nad Váhom
Ul. Sládkovičova 533/20
018 41 Dubnica nad Váhom
E mail: ladislav.mura@gmail.com

Článok je súčasťou riešenia čiastkovej etapy inštitucionálneho výskumu s názvom „Internacionalizácia podnikateľskej činnosti malých a stredných podnikov vo vybranom samosprávnom kraji“, ktorý sa rieši na Ústave odborných predmetov a informačných technológií Dubnického technologického inštitútu v Dubnici nad Váhom.

Východiska prognózy hrubého domácího produktu let 2015 a 2020
Resources of the forecast for gross domestic product in 2015 and 2020

Petr Musil, Jaroslav Zbranek

Abstract: The aim of this article is to describe methods and data sources used for GDP forecast in 2015 and 2020. This forecast will be used for an outlook of the environment. Prognosis based on the development in past only may be insufficient because of changes in economy. This forecast is based on predictable changes and development of each component of GDP. On the other hand a past development is also taken into account.

Key words: GDP, forecast, national accounts

Klíčová slova: HDP, výhled, národní účetnictví

JEL classification: E

1. Úvod

Prognóza úrovně hrubého domácího produktu je poměrně běžnou záležitostí každé země, odlišnosti mohou mezi zeměmi v tomto ohledu spočívat zejména v horizontu prognózy. Jak už název napovídá, cílem článku není prezentovat přímo hodnoty představující odhadnutou úroveň hrubého domácího produktu v uvedeném časovém horizontu. Vzhledem k tomu, že se jedná o časový horizont nezvykle dlouhý v podmínkách České republiky, je vhodné nejprve shromáždit relevantní datové zdroje a další využitelné informace, které budou pro následnou prognózu využity. Obzvláště v současné době, pozvolného oživování po hospodářské krizi, je zaměření pozornosti sběr podstatných relevantních informací stěžejní pro docílení kvalitní prognózy hrubého domácího produktu České republiky v roce 2015 a 2020.

2. Informační zdroje využitelné pro prognózu hrubého domácího produktu

K prognóze hrubého domácího produktu (HDP) budou z velké části využita data z národních účtů, resp. z tabulek dodávek a užití, které každoročně publikuje Český statistický úřad (ČSÚ). Tento systém nabízí možnosti, kterých lze při prognóze jednotlivých složek HDP využít. Tabulky dodávek a užití jsou publikovány v obvyklé formě, kdy produkce a mezispotřeba jsou uvedeny maticově v členění odvětvovém a komoditním. Cílem není prognóza složek HDP v komoditním členění, z toho důvodu bude toto členění použito pouze jako určitý pomocný analytický nástroj. Další zjednodušení pak bude představovat agregace odvětví na několik vytipovaných skupin.

K dosažení stanoveného cíle je třeba vzít v potaz rovněž makroekonomických prognóz Ministerstva financí ČR, byť se jedná pouze o omezený zdroj informací, jelikož tyto informace jsou pouze pro tříletý časový horizont. Na druhou stranu představují tyto informace velmi sofistikovaný základ, zejména pak v současných politických podmínkách.

Vzhledem k tomu, že ČR představuje v současnosti z makroekonomického pohledu malou otevřenou ekonomiku, která je pod silným vlivem zahraničního obchodu, důležitou roli bude hrát také prognóza čistého exportu. V tomto kontextu jsou velmi významné informace resp. prognózy německého hospodářství, protože české národní hospodářství je na německém silně závislé, právě díky zahraničnímu obchodu. Rovněž Slovensko představuje z hlediska zahraničního obchodu významného partnera. Proto i vývoj Slovenského národního hospodářství je směrodatný pro vývoj českého HDP.

Velmi důležité výchozí informace představují rovněž dlouhodobé plány konkrétních subjektů soukromé, ale i veřejné sféry, které mají významný vliv na národní hospodářství. V tomto ohledu jsou zajímavé informace o plánovaných investičních akcích.

3. Předpokládaný vývoj složek hrubého domácího produktu

Složkami HDP se v tomto článku rozumí jednotlivé složky užití. HDP může být obecně užit jednak při další výrobě, ale hlavně pro konečnou spotřebu a dále pro tvorbu kapitálu, resp. k investování. Konečnou spotřebou se rozumí výdaje domácností na individuální konečnou spotřebu, případně výdaje vládních institucí či neziskových institucí sloužících domácnostem na konečnou spotřebu. Zejména pro vládní instituce jsou typické výdaje na kolektivní spotřebu, tedy tu část finální produkce, kterou spotřebovávají všichni rezidenti i nerezidenti daného státu, kteří na území daného státu pobývají. Mezi statky kolektivní spotřeby se řadí např. výdaje ve školství, zdravotnictví, na obranu apod. Specifickou složku HDP představuje čistý vývoz, který představuje rozdíl mezi vývozem a dovozem, kde vývoz je jednou ze složek užití HDP a dovoz pak zdrojů.

V následujících tabulkách je shrnutí vývoje HDP a jeho složek od roku 2000 do roku 2009.

Tabulka 1: Vývoj jednotlivých složek HDP (%)

	2007	2008	2009	průměr 2000 až 2009
HDP	6,1	2,5	-4,1	3,3
Výdaje na konečnou spotřebu	3,6	2,8	0,6	2,8
z toho domácností	4,9	3,6	-0,3	3,0
z toho vládních institucí	0,5	1,1	2,6	2,2
z toho NISD	12,8	5,3	7,9	3,8
Hrubá tvorba kapitálu	9,4	-2,8	-15,8	2,6
z toho hrubá tvorba fixního kapitálu	10,8	-1,5	-7,9	2,9
Vývoz	15,0	6,0	-10,8	9,2
Dovoz	14,3	4,7	-10,6	8,4

Tabulka 2: Příspěvky jednotlivých složek k růstu

	2007	2008	2009	průměr 2000 až 2009
HDP	6,1	2,5	-4,2	3,3
Výdaje na konečnou spotřebu	2,5	1,9	0,4	2,0
z toho domácností	2,3	1,7	-0,2	1,5
z toho vládních institucí	0,1	0,2	0,5	0,5
z toho NISD	0,1	0,0	0,1	0,0
Hrubá tvorba kapitálu	2,5	-0,8	-4,0	0,7
z toho hrubá tvorba fixního kapitálu	2,7	-0,4	-1,9	0,8
Čistý vývoz	1,1	1,3	-0,6	0,6

Výdaje na konečnou spotřebu domácností

Výdaje na konečnou spotřebu domácností (VKSD) tvoří největší složku HDP podle výdajové metody. V první dekádě rostly VKSD v průměru o 3 % ročně, tj. obdobně jako HDP. Domácnosti se tedy podílely na růstu HDP růstem své životní úrovně. Uprostřed dekády ovšem rostl objem úvěrů poskytnutých sektoru domácností až k hodnotám přes 100 000 mil. korun (maximum 261 103 mil. Kč v roce 2007). Většina půjček je dlouhodobých, pravděpodobně převažují půjčky na pořízení nemovitostí (hypoteční úvěry apod.). Investice do nemovitostí patří k jediným výdajům domácností, které nepatří do výdajů na konečnou spotřebu domácností, ale do položky tvorba hrubého fixního kapitálu. Určit přesně VKSD, jejichž zdrojem financování byly úvěry, není možné.

Přestože v roce 2009 VKSD reálně klesly o 0,3 %, pro tvorbu výhledů předpokládáme mírný růst z důvodu odeznívající recese a stabilizace na trhu práce. Nicméně růst bude pouze pozvolný, protože příjmy domácností porostou reálně pomalu z důvodu snížení objemu náhrad zaměstnancům ve veřejném sektoru, míry nezaměstnanosti a obezřetnosti domácností. Na chování domácností budou mít vliv i administrativní změny a to především změny nepřímých daní a to zejména zvýšení snížené sazby DPH. Ačkoliv ve snížené sazbě DPH se nacházejí pouze potraviny, léky, většina nemovitostí určených k bydlení a některé služby, jedná se o sociálně citlivé položky. Domácnosti pravděpodobně příliš neomezí spotřebu těchto základních komodit, ale mohou omezit spotřebu luxusních komodit (substituční efekt). V případě nákupu nemovitosti, které patří do HTFK, by zvýšení sazby DPH mělo vliv na jednorázové zvýšení poptávky v období bezprostředně předcházejícím zvýšení sazby daně. Naopak v období těsně po zvýšení daně dojde ke krátkodobému poklesu poptávky, ale vliv v horizontu výhledů nebude pravděpodobně významný. Jelikož zatím nejsou známy detaily, především nová sazba daně a její platnost, je obtížné kvantifikovat vliv změny daní na VKSD.

Dalším faktorem pro modelování výdajů na konečnou spotřebu domácností v národním pojetí (tj. výdaje rezidentů v ČR i v zahraničí) je počet obyvatel. Demografické prognózy se sestavují ve třech variantách (nízká, střední, vysoká), a proto pro každou variantu vývoje HDP využijeme příslušnou demografickou prognózu.

Výdaje konečnou spotřebu vládních a neziskových institucí

Podle definice standardu ESA 95 se výdaje na konečnou spotřebu vládních institucí rovnají netržní produkci těchto institucí a naturálním sociálním dávkám. Netržní produkce je odhadována jako součet nákladů (mezispotřeby, náhrad zaměstnancům, spotřeby fixního kapitálu a čistým daním na produkci). Pro příští rok (2011) je snížen objem mzdových prostředků v nepodnikatelské sféře o 10 % se zafixováním na několik dalších let a snížení provozních výdajů. Tato opatření budou mít vliv na velikost netržní produkce (snížení), spotřeba fixního kapitálu je modelována bez ohledu na účetní odpisy. Naopak u naturálních sociálních dávek (produkty vyrobené tržními výrobci poskytované domácnostem za ekonomicky nevýznamné ceny např. léky) neočekáváme snížení. Obecně vývoj tohoto makroagregátu závisí na politických rozhodnutích, která závisejí na výsledku voleb.

Výdaje neziskových institucí sloužících domácnostem tvoří velmi malou část HDP, a proto se jí ve výhledech nebudeme detailně zabývat.

Hrubá tvorba kapitálu

Jak již bylo uvedeno, hrubá tvorba kapitálu (HTK) představuje složku užití HDP, která v sobě zahrnuje zejména investice na národohospodářské úrovni nazývané jako hrubá tvorba fixního kapitálu (HTFK). Makroekonomický agregát HTK v sobě dále zahrnuje změnu stavu zásob a čisté pořízení cenností. Nicméně změna stavu zásob je v dlouhodobém horizontu

nulová, neboť jak už samotný název napovídá, může nabývat jak kladných tak záporných hodnot. Proto změnu stavu zásob do prognózy HDP nikterak detailněji nezahrnujeme. Pokud jde o čisté pořízení cenností, tyto představují dlouhodobě poměrně zanedbatelný podíl jak na HDP tak na HTK, průměrný podíl představuje přibližně 0,1 % resp. 0,4 %. Z toho důvodu také se tímto makroagregátem při tvorbě prognózy rovněž blíže zabývat nebudeme. Z uvedeného vyplývá, že jediná složka HTK, na kterou bude při prognóze zaměřena pozornost, je HTFK.

Makroekonomický agregát THFK lze díky své povaze jen velmi obtížně s větší spolehlivostí predikovat. Na základě dlouhodobých plánů významných tržních i netržních subjektů lze usuzovat, alespoň přibližně, zda bude investiční aktivita významnější než v současnosti, či naopak.

Mezi takové subjekty patří např. energetická společnost ČEZ, a.s., která pro uvedené období plánuje dostavbu 3. a 4. bloku jaderné elektrárny Temelín. Vláda se zmiňovanou dostavbou díky očekávané poptávce po elektrické energii příliš nepočítá, avšak samotná energetická společnost tuto investici stále předpokládá, i když s odloženou realizací. Rovněž ŠKODA AUTO, a.s. během horizontu prognózy plánuje investiční projekty nejen v souvislosti s modernizací, ale také s rozšířením výrobních možností, resp. s rozšířením produktové nabídky.

Další fakt, který by měl být do prognóz zahrnut, je nutnost produkce elektrické energie z obnovitelných zdrojů, aby spotřeba takové energie představovala 13 % celkové spotřeby energie. Přitom v roce 2009 představoval podíl spotřeby elektrické energie z obnovitelných zdrojů na celkové spotřebě pouze 6,8 %. Z této skutečnosti vyplývá nutnost investic do výstavby příslušných zařízení pro výrobu potřebného množství elektrické energie. Blíže k odhadům ukazatelů z oblasti energetiky v [2].

Pokud jde o investice subjektů sektoru vládních institucí, lze očekávat poměrně významné investiční akce. S ohledem na současnou situaci v tomto sektoru je však zřejmě nezbytné počítat s rozložením investičních projektů do delšího časového intervalu, případně s odložením jejich začátku. Mezi takové projekty lze řadit např. investice spočívající ve výstavbě a zejména pak v rekonstrukci silniční sítě. Jiným ožehavým příkladem nutných investičních výdajů jsou nápravná opatření související s delikty vůči životnímu prostředí s dob minulých.

Čistý vývoz

Česká ekonomika patří mezi malé otevřené ekonomiky s vysokým podílem vývozu na HDP, který v první dekádě stoupal až do roku 2007, kdy dosáhl 80 %. Vývoz z České republiky závisí především na ekonomickém vývoji v ostatních zemích Evropské unie, kam směřuje většina vývozu. Dovoz do ČR tvoří produkty, které nelze v ČR vyrobit (například ropa). Druhou skupinou je dovoz produktů, které slouží jako vstupy pro výrobu zboží, které je následně vyvezeno (například automobilové díly, procesory do počítačů) nebo jde o technologie pro tento typ výroby. Dovoz tedy částečně závisí i na vývozu.

Pro přibližnou kvantifikaci závislosti mezi vývozem a ekonomickým vývojem v zahraničí je možné využít korelační analýzu. Ve sledovaném období je korelační koeficient mezi reálným vývojem vývozu a HDP v EU 27 0,92. Obdobný výsledek (0,89) vyjde při měření závislosti mezi vývozem a vývojem HDP v Německu, které je největším obchodním partnerem ČR. Je tedy zřejmé, že vývoj čistého vývozu v dalším období bude záviset na ekonomickém vývoji v Evropské unii. Modelování vývoje HDP v EU je výrazně obtížnější, protože recese zasáhla jednotlivé země různě a různé jsou i jejich vyhlídky pro nadcházející období. Ekonomický vývoj v EU nahradíme vývojem v Německu, které je naším nejdůležitějším obchodním partnerem.

4. Závěr

Prognózování vývoje HDP coby jednoho z nejvýznamnějších ekonomických ukazatelů národního hospodářství je velmi komplikovaná činnost. Faktorů ovlivňujících vývoj HDP je velmi mnoho a některé jsou i velmi obtížně kvantifikovatelné. Pro výhledy byla vybrána výdajová metoda, neboť je možné alespoň v omezené míře analyzovat faktory, které působí na jednotlivé složky HDP, které se mohou v čase vyvíjet odlišně. Mnohé makroagregáty, které představují složky užití HDP, se řídí určitými obecnými zákonitostmi. Tyto skutečnosti následně vedou do jisté míry k přesnějším odhadům samotného HDP, lépe řečeno k redukci chyby odhadu.

5. Literatura

[1] HRONOVÁ, S., FISCHER, J., HINDLS, R., SIXTA, J. 2009. Národní účetnictví. Nástroj popisu globální ekonomiky. 2009, ISBN 978-80-7400-153-6.

[2] JEDLIČKOVÁ, E., KONVIČKA, S., MUSIL, P., SIXTA, J. 2009. Nové metody deflace HDP. Statistika Roč. 89, č. 3, 193--206.

[3] Evropský systém účtů (ESA 1995). ČSÚ, Praha 2000.

[4] www.czso.cz

[5] www.mfcr.cz

[6] www.cez.cz

[7] www.skoda-auto.cz

Adresa autora (-ov):

Petr Musil, Ing.
VŠE Praha
Katedra ekonomické statistiky
Nám. W. Churchilla 4
130 67 Praha 3
petr.musil@vse.cz

Jaroslav Zbranek, Ing.
VŠE Praha
Katedra ekonomické statistiky
Nám. W. Churchilla 4
130 67 Praha 3
jaroslav.zbranek@vse.cz

Tento text vznikl s podporou grantu GA ČR č. 402/10/1275 "Historické časové řady HDP ČR"

Poznámka k asociačným koeficientom
One expression of associative coefficients

Nánásiová Oľga, Pastuchová Elena, Václavíková Štefánia

Abstract: In engineering practice are used many different association coefficients to compare the similarity of investigated objects. In the present paper some of associative coefficients are classified by a type of a rational function by using of symmetric difference.

Abstract: Asociačné koeficienty sú v inžinierskej praxi často používané na vyjadrenie podobnosti sledovaných objektov.V článku sú analyzované vybrané typy koeficientov pomocou symetrickej diferencie.

Key words: associative coefficients, similarity, dissimilarity, symmetric difference

Key words: asociačné koeficienty, podobnosť, nepodobnosť, symmetrická diferencia

JEL classification: C 38

1. Introduction

Studies comparing similarity coefficient are common, but each study has generated different conclusions, prompting a general acceptance that the "behavior "of similarity coefficients is data specific. As a result, choice of a similarity coefficients is largely subjective and often based on tradition or on posteriori criteria such as the "interpretability"of the results.[4]

This and similar conclusions led us to the idea, why some of this coefficient give identical results and other not? Is it really "subjective" depends on researcher`s choice?

Our ambition was find out the entity of this problem.Our results and conclusions of the articles [1], [4],[8],[9] confirm us, that Jaccard's, Sorensen-Dice's coefficients and other showed close results, due to the fact, that all of them exclude negative co-occurrence.

We identify a function, rational function, that helps us to divide some of known similarity coefficients into 2 categories :

1.Dice's, Jaccard's, Sokal-Sneath's(2) and other

2.Sokal-Sneath's (1), Rogerr-Tanimoto's and other.

2. Measure of similarity and dissimilarity

In this part we define basic notions and its properties. We start with definitions of similarity and dissimilarity of two objects.

Definition 1. [1] Let E be a not empty set. A function $s: E^2 \to \langle 0,1 \rangle$ is called a measure of similarity if the following statements are hold:

1. $s(x,y) = s(y,x)$ for $x, y \in E$.
2. $s(x,x) = 1$ for each $x \in E$.

Definition 2. [1] Let E be a not empty set. A function $s: E^2 \to \langle 0,1 \rangle$ is called a measure of dissimilarity if the following statements are hold:

1. $d(x,y) = d(y,x)$ for $x, y \in E$.
2. $d(x,x) = 0$ for each $x \in E$.

It is easy to see that for any similarity measure there exists a dissimilarity measure and conversely. Apparently duality can be expressed as follows: $s(x,y)=1-d(x,y)$ for all $x,y\in E$.

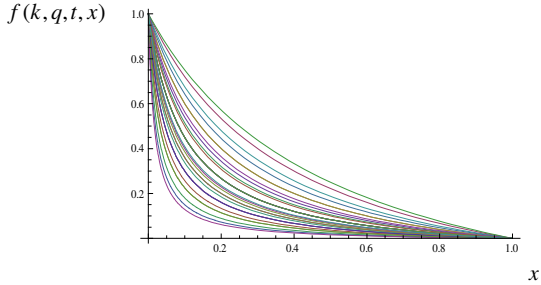
3. Rational functions and associative coefficient

Let $k>0,\quad q\geq 0$ and $x\in\langle 0,1\rangle$. Then mentioned rational function is:

$$f(k,q,t,x)=\frac{k(1-x)}{t+qx}. \tag{1}$$

Clearly $f(k,q,t,1)=0$ and $f(k,q,t,0)=\frac{k}{t}$.

Let $k=t,$ and $x_1<x_2$ then $f(k,q,x_1)>f(k,q,x_2)$. It means that, it has good properties for dissimilarity function.



Picture 1: Graph of rational functions (1)

Let (Ω,S,P) be a measurable space with a normed measure. Let functions $s:S^2\to R$ and $g:S^2\to R$ be defined in the following way:

$$s(A,B)=\frac{P(A\cap B)}{P(A\cap B)+\alpha P(A\cap B^C)+\beta P(B\cap A^C)} \tag{2}$$

$$g(A,B)=\frac{P(A\cap B)+\gamma P(A^C\cap B^C)}{P(A\cap B)+\alpha P(A\cap B^C)+\beta P(B\cap A^C)} \tag{3}$$

It is apparent, that $s(A,A)=g(A,A)=1$ for all $A\in S$, such that $P(A)>0$.

If in (2) $\alpha=\beta,$ then

$s(A,B)=\frac{P(A\cap B)}{P(A\cap B)+\alpha P(A\Delta B)}=\frac{P(A\cup B)-P(A\Delta B)}{P(A\cup B)-P(A\Delta B)+\alpha P(A\Delta B)}$, when operator $(A\Delta B)=(A\cap B^C)+(A^C\cap B)$

It is easy to see, that

$$s(A,B)=\frac{P(A\cup B)-P(A\Delta B)}{P(A\cup B)+(\alpha-1)P(A\Delta B)} \tag{4}$$

Let $P(A \cup B) > 0$. Therefore $A \Delta B \subseteq A \cup B$, then $\frac{P(A \Delta B)}{P(A \cup B)} = P(A \Delta B \mid P(A \cup B)$

$$s(A,B) = \frac{1 - P(A \Delta B \mid A \cup B)}{1 - (\alpha - 1) P(A \Delta B \mid A \cup B)} \tag{5}$$

Let's assume, that $P(A \cup B) \in (0,1)$ and let us put $P(A \Delta B \mid A \cup B) = x$. Then the function s can be revrite by the following way:

$$s(A,B) = \frac{1-x}{1-(\alpha-1)x} \tag{6}$$

We can see, that $s(A,B) = \frac{1-x}{1-(\alpha-1)x} = f(1, \alpha - 1, 1, x)$ is a special form of (1).

4. Group G1

In like manner, if $P(A \Delta B \mid A \cup B) = x$, each of known similarity coefficient from the set $G1$ can be express as a special form of this function, where $k, t, q \in R$.

1. Dice`s coefficient: $s_1(A,B) = \frac{2(P(A \cup B) - P(A \Delta B))}{2P(A \cup B) - P(A \Delta B)}$ can be rewrite as

$$s_1(A,B) = \frac{2(1-x)}{2-x} = f(2,-1,2,x) \ ,$$

where $P(A \Delta B \mid A \cup B) = x$.

2. Jaccard`s coefficient: $s_2(A,B) = 1 - \frac{P(A \Delta B)}{P(A \cup B)}$

$$s_2(A,B) = 1 - x = f(1,0,0,x),$$

where $P(A \Delta B \mid A \cup B) = x$.

3. Sokal –Sneath`s coefficient (2) : $s_3(A,B) = \frac{P(A \cup B) - P(A \Delta B)}{P(A \cup B) + P(A \Delta B)}$

$$s_3(A,B) = \frac{1-x}{1+x} = f(1,1,1,x),$$

where $P(A \Delta B \mid A \cup B) = x$.

5. Group G2

On the other hand, coefficients from $G2$ can be express using $x = P(A \Delta B)$ as

4. Rogerr-Tanimoto`s coefficient: $s_4(A,B) = \frac{1 - P(A \Delta B)}{1 + P(A \Delta B)}$

$$s_4(A,B) = \frac{1-x}{1+x} = f(1,1,1,x),$$

where $x = P(A \Delta B)$.

5. Sokal –Sneath`s coefficient (1): $s_5(A,B) = \frac{2(1 - P(A \Delta B))}{2 - P(A \Delta B)}$

$$s_5(A,B) = \frac{2(1-x)}{2-x} = f(2,-1,2,x),$$

where $x = P(A \Delta B)$.

6. Conclusion

We can see that coeficients s_1, s_5 have the same properties as the hyperbolic function $f(2,-1,2,x)$ and s_3, s_4 have the same properties as the hyperbolic function $f(1,1,1,x)$, but coeficient s_2 is defined by the linear funcion $f(1,0,0,x) = 1-x$. While the coefficients from $G1$ compare elements inside this two sets, only with respect to $A \cup B$, ergo substitution ($x = P(A \Delta B \mid A \cup B) \in [0,1]$), coefficient from $G2$ are the measure of similarity with respect to all elements of Ω, ergo $x = P(A \Delta B)$.

7. References

[1] Dalisferat S.B., Mayer A., Mirhoseini S. Z., 2009 : *Journal of Insect Science* Vol.9,71, 681-689

[2] Duarte M.C., Santos J.B., Melo L.C., 1999 : Comparison of similarity coefficients based on RAPD markers in common bean. *Genetic and molecular biology* 22: 427-432

[3] Haldiki M., Batistakis Y., Vazirgiannis M., 2001: On clustering Validation Techniques. *Journal of Intelligent Information System* 17:2/3, 107-145

[4] Jackson, A.D., Somers K.M., & Harvey, H.H. 1989: Similarity coefficients: Measures of co-occurrence and association or simply measures of occurrence. *The American Naturalist* Vol.133, No.3; 436-453.

[5] Johnson RA, Wichern DW. 1988. Applied multivariate statistical analysis. *Prentice-Hall.* 458-465.

[6] Keshavarzi,M., Dehghan,M.A.,Mashinchi,M. 2009: Classification based on similarity and dissimilarity through equivalence classes.-*Appl.and Comput.Math*., Vol.8, No.2.,203-215

[7] Kosman E, Leonard KJ. 2005: Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species. *Molecular Ecology* 14; 415-424.

[8] Murgula, M.& Villasenor, J.L. 2003: Estimating the effect of similarity coefficient and the cluster algorithm on biogeographic classifications. *Ann. Bot. Fennici* 40; 415-421

[9] Sepkoski,J.J.,Jr. 1974*:* Quantified Coefficients of Association and Measurement of Similarity. *Mathematical Geology*, Vol.6, No.2, 135-141.

[10] Si Quang Le, & Tu Bao Ho. 2005: An association-based dissimilarity measure foe categorical data. *Elsevire,*Vol .26, Issue 16, 2549-2557.

Addreses :

Pastuchová Elena, RNDr., PhD
KM FEI STU
Ilkovičova 3, Bratislava
elena.pastuchova@stuba.sk

Václavíková Štefánia, Mgr.
KMaDG SvF STU
Radlinského 11, Bratislava
vaclavikova@is.stuba.sk

Nánásiová Oľga, Doc., RNDr., PhD.
KMaDG SvF STU
Radlinského 11, Bratislava
nanasiova@is.stuba.sk

Aknowlegement: *This work was supported by the Slovak Research and Development Agency under the contract No APVV-0249-07, VEGA 1/0373/08 and VEGA 1/0103/10.*

Detekce změn v českém indexu spotřebních cen pomocí metody hledání řídkých řešení

Detection of Changes in the Czech Consumer Price Index by Sparse Parameter Estimation

Jiří Neubauer, Vítězslav Veselý

Abstract: The contribution is focused on application of multiple change point detection in a one-dimensional stochastic process by sparse parameter estimation from an overparametrized model. A stochastic process with changes in the mean is estimated using dictionary consisting of Heaviside functions. The basis pursuit algorithm is used to get sparse parameter estimates. The mentioned method is applied to the time series of the Czech consumer price index.

Abstrakt: Článek je zaměřen na aplikaci metody detekce vícenásobných změn v jedno-rozměrném stochastickém procesu pomocí hledání řídkých řešení v přeparametrizovaném modelu. Slovník skládající se z Heaviside funkcí je používán pro odhady změn ve střední hodnotě stochastického procesu. Řídké odhady byly získány pomocí basis pursuit algoritmu. Uvedená metoda byla aplikována na časovou řadu českého indexu spotřebitelských cen.

Key words: multiple change point detection, overparametrized model, sparse parameter estimation, basis pursuit algorithm, consumer price index.

Klíčová slova: detekce vícenásobných změn, přeparametrizovaný model, řídká řešení, basis pursuit algoritmus, index spotřebitelských cen.

JEL classification: C21, C51.

1. Introduction

Detection of changes plays an important role in statistical analysis of time series (see Antoch et al. (2000)). An alternative way of change point identification in the one-dimensional stochastic process is to use methods of sparse parameter estimation. This article deals with application of the so-called basis pursuit algorithm with the Heaviside dictionary to the time series of the Czech consumer price index (CPI).

Chen, S. S. et al. (1998) proposed a new methodology based on basis pursuit for spectral representation of signals (vectors). Instead of just representing signals as superpositions of sinusoids (the traditional Fourier representation) they suggested alternate dictionaries – collections of parametrized waveforms – of which the wavelet dictionary is only the best known. A recent review paper by Bruckstein et al. (2009) demonstrates a remarkable progress in the field of sparse modeling since that time. A lot of other useful applications in a variety of problems can be found in Veselý and Tonner (2005) and Veselý et al. (2009). In Zelinka et al. (2004) kernel dictionaries showed to be an effective alternative to traditional kernel smoothing techniques. In this paper we are using Heaviside dictionary in the same manner to denoise signal (the univariate time series sample path) exhibiting jumps (discontinuities) in the mean (unlike kernel smoothing where the mean is supposed to be sufficiently smooth). Consequently, the basis pursuit approach can be proposed as an alternative to conventional statistical techniques of change point detection (see Neubauer and Veselý (2009)). The mentioned paper is focused on application of the basis pursuit algorithm with the Heaviside dictionary to one change point detection. The method of multiple change

point detection and basic properties of this approach to identification of changes in one-dimensional processes are described in Neubauer and Veselý (2010).

This paper presents results of one possible application of the proposed method of multiple change point detection. We will focus on time series of the Czech consumer price index (monthly data 2000–2010) and we will show how to detect possible changes in the mean by the method of sparse parameter estimation.

2. Heaviside Dictionary for Change Point Detection

In this section we propose the method based on basis pursuit algorithm (BPA) for the detection of the change point in the sample path $\{y_t\}$ in one dimensional stochastic process $\{Y_t\}$. We assume a deterministic functional model on a bounded interval I described by the dictionary $G = \{G_j\}_{j\in J}$ with atoms $G_j \in L^2(I)$ and with additive white noise e on a suitable finite discrete mesh $\tilde{T} \subset I$:

$$Y_t = x_t + e_t, t \in \tilde{T},$$

where $x \in sp(\{G_j\}_{j\in J})$, $\{e_t\}_{t\in\tilde{T}} \sim \mathrm{WN}(0,\sigma^2)$, $\sigma > 0$, and J is a big finite indexing set. Smoothed function $\hat{x} = \sum_{j\in J} \hat{\xi}_j G_j =: \mathbf{G}\hat{\xi}$ minimizes on $\tilde{T}$ ℓ^1-penalized optimality measure $\frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2$ as follows:

$$\hat{\xi} = \operatorname{argmin}_{\xi\in\ell^2(J)} \frac{1}{2}\|\mathbf{y} - \mathbf{G}\xi\|^2 + \lambda\|\xi\|_1, \|\xi\|_1 := \sum_{j\in J} \|G_j\|_2 \, \xi_j,$$

where $\lambda = \sigma\sqrt{2\ln(\operatorname{card} J)}$ is a smoothing parameter chosen according to the soft-thresholding rule commonly used in wavelet theory. This choice is natural because one can prove that with any orthonormal basis $G = \{G_j\}_{j\in J}$ the shrinkage via soft-thresholding produces the same smoothing result $\hat{x}$. (see Bruckstein et al. (2009)). Such approaches are also known as basis pursuit denoising (BPDN).

Solution of this minimization problem with λ close to zero may not be sparse enough: we are searching small $F \subset J$ such that $\hat{x} \approx \sum_{j\in F} \hat{\xi}_j G_j$ is a good approximation. That is why we apply the following four-step procedure described in Zelinka et al. (2004) in more detail and implemented in Veselý (2001–2008).

(A0) Choice of a raw initial estimate $\xi^{(0)}$, typically $\xi^{(0)} = \mathbf{G}^{+}\mathbf{y}$.

(A1) We improve $\xi^{(0)}$ iteratively by stopping at $\xi^{(1)}$ which satisfies optimality criterion BPDN. The solution $\xi^{(1)}$ is optimal but not sufficiently sparse in general (for small values of λ).

(A2) Starting with $\xi^{(1)}$ we are looking for $\xi^{(2)}$ by BPA which tends to be nearly sparse and is optimal.

(A3) We construct a sparse and optimal solution ξ^{*} by removing negligible parameters and corresponding atoms from the model, namely those satisfying $|\xi_j^{(2)}| < \alpha\|\xi^{(2)}\|_1$ where $0 < \alpha << 1$ is a suitable sparsity level, a typical choice being $\alpha = 0.05$ following an analogy with the statistical significance level.

(A4) We repeat the step (A1) with the dictionary reduced according to the step (A3) and with a new initial estimate $\xi^{(0)} = \xi^{*}$. We expect to obtain a possibly improved sparse

estimate ξ^*.

Hereafter we refer to this four-step algorithm as to BPA4. The steps (A1), (A2) and (A4) use Primal-Dual Barrier Method designed by M. Saunders (see Saunders (1997–2001)). This up-to-date sophisticated algorithm allows one to solve fairly general optimization problems minimizing convex objective subject to linear constraints. A lot of controls provide a flexible tool for adjusting the iteration process.

We build our dictionary from heaviside-shaped atoms on $L^2(\mathbf{R})$ derived from a fixed 'mother function' via shifting and scaling following the analogy with the construction of wavelet bases.

We construct an oversized shift-scale dictionary $G = \{G_{a,b}\}_{a\in A, b\in B}$ derived from the 'mother function' by varying the shift parameter a and the scale (width) parameter b between values from big finite sets $A \subset \mathbf{R}$ and $B \subset \mathbf{R}^+$, respectively ($J = A \times B$), on a bounded interval $I \subset R$ spanning the space $H = \mathrm{sp}(\{G_{a,b}\})_{a\in A, b\in B}$, where

$$G_{a,b}(t) = \begin{cases} 1 & \text{for} \quad t - a > b/2, \\ 2(t-a)/b & \quad |t-a| \le b/2, b > 0, \\ 0 & \quad t = a, b = 0, \\ -1 & \quad \text{otherwise.} \end{cases}$$

Some examples of Heaviside functions are displayed in the figure 1.

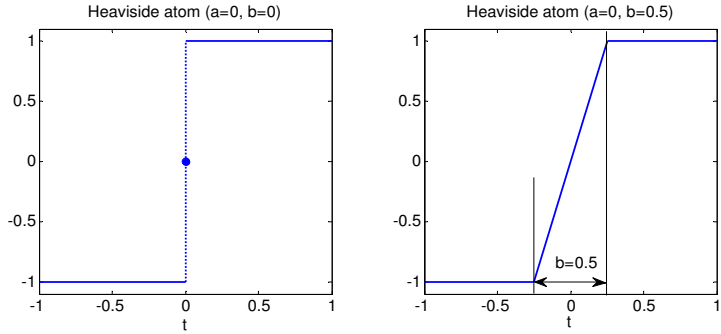


Figure 7: Heaviside atoms with parameters a = 0, b = 0 and a = 0, b = 0.5

3. Change Point Detection by Basis Pursuit

Neubauer and Veselý (2009) proposed the method of change point detection if there is just one change point in a one-dimensional stochastic process (or in its sample path). We briefly describe the given method. We would like to find a change point in a stochastic process

$$Y_t = \begin{cases} \mu + \varepsilon_t & t = 1, 2, \ldots, c, \\ \mu + \delta + \varepsilon_t & t = c+1, \ldots, T, \end{cases} \tag{1}$$

where $\mu, \delta \neq 0, t_0 \leq c < T - t_0$ are unknown parameters and ε_t are independent identically distributed random variables with zero mean and variance σ^2 (t_0 is a boundary trimming). The parameter c indicates the change point in the process. Using BPA4 we obtain some significant atoms and calculate correlation between significant atoms and the analyzed process. The shift parameter of the atom with the highest correlation is taken as an estimator of the change point c.

4. Multiple Change Point Detection by Basis Pursuit

Now let us assume the model with *p* change points

$$Y_t = \begin{cases} \mu + \varepsilon_t & t = 1, 2, \ldots, c_1, \\ \mu + \delta_1 + \varepsilon_t & t = c_1 + 1, \ldots, c_2, \\ \ldots & \ldots \\ \mu + \delta_p + \varepsilon_t & t = c_p + 1, \ldots, T, \end{cases} \tag{2}$$

where $\mu, \delta_1, \delta_2, \ldots, \delta_p \neq 0, t_0 \leq c_1 < c_2 < \cdots < c_p < T - t_0$ are unknown parameters and ε_t are independent identically distributed random variables with zero mean and variance σ^2.

We use a modified method of change point estimation described above for detection of *p* change points $c_1, c_2, \ldots, c_p$ in the model (2). Instead of finding only one significant atom with the highest correlation with the process Y_t we can identify *p* significant atoms with the highest correlation. The shift parameters of these atoms determine estimators of the change points c_1, $c_2, \ldots, c_p$. Another possibility is to apply the procedure of one change point detection *p* times in sequence. In the first step we identify one change point in the process Y_t and subtract the appropriate significant atom from the process (by linear regression)

$$Y_t = \beta G_{0,\hat{c}_1} + e_t,$$
$$Y_t' = Y_t - \hat{\beta} G_{0,\hat{c}_1}.$$

Then we apply the method to the new process Y_t' and repeat this procedure until getting all *p* estimates. The shift parameters of selected atoms are again identifiers of the change points $c_1, c_2, \ldots, c_p$. Observe that this can be seen as *p*-steps of orthogonal matching pursuit (OMP, see Bruckstein (2009)) combined with BPA.

We will analyze the time series of the Czech consumer price index (monthly data 2000–2010) applying the method of multiple change point detection described above. The data set of this time series comes from the database of the Czech National Bank and can be downloaded from the webpage http://www.cnb.cz/docs/ARADY/HTML/index.htm. It consists of 129 monthly measured values of CPI from 31. 1. 2000 to 30. 9. 2010. The basis of the index (level 100) corresponds to the same period in the preceding year. Using both variants of the mentioned method we get the same result indicating four significant change points 29, 48, 93 and 108 which means that there were some changes in the behavior of CPI in May 2002–June 2002, December 2003–January 2004, September 2007–October 2007 and December 2008–January 2009 (see figure 2). When striving after the reasons for these change points, we can have a look at the webpage of the Czech Statistical Office, the section 'analysts' (http://czso.cz/csu/redakce.nsf/i/analyzy_csu). For example, it can be found that the change of CPI in the first quarter of 2004 was caused by an increase of the value added tax (VAT) and the consumer tax, prices of electricity and gas. The change in the 4th quarter of

2007 was brought about by the increase of the consumer tax and the increase of natural gas. The CPI decrease early in 2009 (compared to the year 2008) originated from the price fall of driving fuel and passenger vehicles.

5. Conclusion

This article briefly describes one possible approach to multiple change point detection which seems to be a reasonable alternative to classical statistical methods. We applied this method to the time series of Czech CPI and indentified the most significant changes in the mean of this time series. The outlined method can be used for detection of two or more change points, or another sort of change point with a dictionary G of different kind which will be the prospective aim of our research.

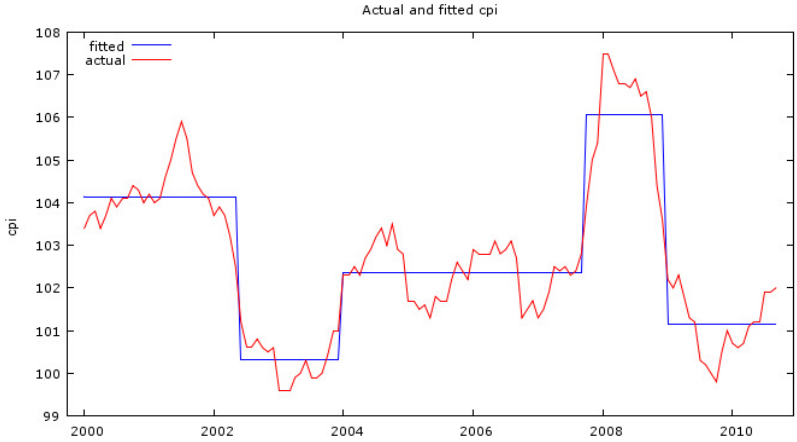


Figure 2: Czech Consumer price index

6. References

[1] ANTOCH, J. – HUŠKOVÁ, M. – JARUŠKOVÁ, D. 2000. Change point detection. In *5th ERS IASC Summer School*, IASC 2000.

[2] BRUCKSTEIN, A. M. et al. 2009. From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images. SIAM Review 51 (1), p. 34–81.

[3] CHEN, S. S. et al. 1998. Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. 20 (1), p. 33–61 (2001 reprinted in SIAM Review 43 (1), p. 129–159).

[4] NEUBAUER, J. – VESELÝ, V. 2009. Change Point Detection by Sparse Parameter Estimation. In: The XIIIth International conference: Applied Stochastic Models and Data Analysis. Vilnius, 158–162.

[5] NEUBAUER, J. – VESELÝ, V. 2010. Multiple Change Point Detection by Sparse Parameter Estimation. In Proceedings of COMPSTAT'2010, p. 1453–1460.

[6] SAUNDERS, M. A. 1997–2001. pdsco.m: MATLAB code for minimizing convex separable objective functions subject to $Ax = b, x \geq 0$.

[7] VESELÝ, V. 2001–2008. framebox: MATLAB toolbox for overcomplete modeling and sparse parameter estimation.

[8] VESELÝ, V. – TONNER, J. 2005. Sparse parameter estimation in overcomplete time series models. Austrian Journal of Statistics, 35 (2&3), p. 371–378.

[9] VESELÝ, V. et al. 2009. Analysis of PM10 air pollution in Brno based on generalized linear model with strongly rank-deficient design matrix. Environmetrics, 20 (6), p. 676–698.

[10] ZELINKA, J. et al. 2004. Comparative study of two kernel smoothing techniques. In: Horová, I. (ed.) Proceedings of the summer school DATASTAT'2003, Svratka. Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica 15: Masaryk University, Brno, Czech Rep., p. 419–436.

Authors' addresses:

Jiří Neubauer, Mgr, Ph.D.
University of Defence
Kounicova 65
662 10 Brno, Czech Republic
Jiri.Neubauer@unob.cz

Vítězslav Veselý, doc., RNDr., CSc.
Masaryk University
Lipová 41a
602 00 Brno, Czech Republic
vesely@econ.muni.cz

Acknowledgment: *supported by grants GAČR P402/10/P209 and MSM0021622418.*

Modelovanie závislosti rezíduí modelu SETAR s použitím autokopúl
Modelling of dependence structure of the SETAR models residuals using autocopulas

Anna Petričková, Jana Lenčuchová

Abstract: The autocorrelation function describing linear dependence is not suitable for description of residual dependence of regime-switching models. Therefore we want to investigate description of this dependence with 'k-lag auto-copula', which is 2-dimensional joint distribution function of the bivariate random vector (Y_t , Y_{t-k}) of time lagged values of random variables that generate time series (in the analogy of the autocorrelation function of stationary time series). In this contribution we will describe the dependence of time lagged residuals of SETAR models by means of copulas, we will test the independence of these residuals and propose „improved" model SETAR.

Abstrakt: Autokorelačná funkcia popisujúca lineárnu závislosť nie je vhodná na popis závislosti rezíduí modelov s premenlivými režimami. Preto chceme vyšetriť popis tejto závislosti použitím 'autokopúl', čo je dvojrozmerná združená distribučná funkcia náhodného vektora (Y_t , Y_{t-k}) časovo posunutých náhodných premenných generujúcich časový rad. V tomto príspevku popíšeme závislosť časovo posunutých rezíduí modelu SETAR pomocou autokopúl, budeme testovať nezávislosť týchto rezíduí a navrhneme „vylepšený" SETAR model.

Key words: time series, regime-switching models, k-lag auto-copula, residuals, BDS test.

Kľúčové slová: časové rady, regime-switching models, autokopula, rezíduá, BDS test.

1. Introduction

The autocorrelation function is suitable for description of residual dependence only in the case of linear models. So the autocorrelation function is not suitable for description of residual dependence of regime-switching models (because these models have nonlinear character). Therefore we investigate description of this dependence with '*k*-lag auto-copula', which is 2-dimensional joint distribution function of the bivariate random vector (Y_t , Y_{t-k}) of time lagged values of random variables that generate time series (in the analogy of the autocorrelation function of linear stationary time series).

At first we have to test independence in residuals $\{\hat{e}_t\}$ by the BDS test. When the BDS test shows residual dependence at the significant level, inspired by the approach of Rakonczai in [6], we applied 2-lag auto-copulas to the time series of the above mentioned residuals and succeeded to construct improved quality models for the original data.

The paper is organized as follows. After a general introduction, the theoretical basis of model SETAR, copulas and some tests are described. The paper continues with their application to a modelling dependence of residuals of real time series with auto-copulas and constructing of improved quality models for the original data.

2. Theoretical basis

Model SETAR

Typical models belonging to the class of regime-switching models, which are well to interpret and are also very suitable for modeling a lot of real data, are TAR models

(„**T**hreshold **A**uto**R**egressive“). They form the basis of regime-switching models with regimes determined by observable variables.

These models assume that any regime in time t can be given by any observed variable q_t (*indicator variable*). Values of q_t are compared with *threshold value* c. In case of 2-regimes model the first regime applies if $q_t \le c$, the second if $q_t > c$.

We have the model SETAR when the variable q_t is taken to be a lagged value of the time series itself, that is $q_t = X_{t-d}$ for a certain integer $d > 0$. The resulting model is called a **S**elf-**E**xciting **T**hreshold **A**uto**R**egressive (SETAR) model. For example 2-regime model SETAR with AR(p) in both regimes has form

$$\begin{aligned} X_t = &\left(\phi_{0,1} + \phi_{1,1}X_{t-1} + \cdots + \phi_{p,1}X_{t-p}\right)\left[1 - \mathbf{1}\left(X_{t-d} > c\right)\right] + \\ &+ \left(\phi_{0,2} + \phi_{1,2}X_{t-1} + \cdots + \phi_{p,2}X_{t-p}\right)\mathbf{1}\left(X_{t-d} > c\right) + e_t \end{aligned} \qquad (1)$$

where $\{e_t\}$ is the strict white noise process with $E[e_t] = 0$, $D[e_t] = \sigma_e^2$ for all t = 1, ..., n and $\mathbf{1}(A)$ is the *indicator function* with values $\mathbf{1}(A) = 1$ if the event A occurs and $\mathbf{1}(A) = 0$ otherwise.

In case of 3-regimes model we have to define 4 constants c_0, c_1, c_2, c_3 where $-\infty = c_0 < c_1 < c_2 \le c_3 = \infty$. Model SETAR with AR(p) in all regimes has form

$$X_t = \phi_{0,j} + \phi_{1,j}X_{t-1} + \ldots + \phi_{p,j}X_{t-p} + e_t \quad \text{if } c_{j-1} < X_{t-d} \le c_j, \quad j = 1, 2, 3 \qquad (2)$$

For more details see [1], [5].

The BDS test

The BDS test presented in [2] can be used to test independence in residuals $\{\hat{e}_t\}$. It is a portmanteau test which have been constructed without a specific nonlinear parametric alternative in mind. This test of independence can be applied to the estimated residuals of any time series model that can be transformed into a model driven by independent and identically distributed errors.

Copulas

2-dimensional copula is a function (see e.g [8])

$$C : [0,1]^2 \to [0,1]$$

such that

$$C(0, y) = C(x, 0) = 0, \quad C(1, y) = y, \; C(x, 1) = x,$$

for all $x, y \in [0, 1]$ and

$$C(x_1, y_1) + C(x_2, y_2) - C(x_1, y_2) - C(x_2, y_1) \ge 0$$

for all $x_1, x_2, y_1, y_2 \in [0, 1]$ with $x_1 \le x_2$, $y_1 \le y_2$.

The most important applications of 2-dimensional copulas are related to a well known, very convenient alternative of expressing of joint distribution function of 2-dimensional random vectors *(X, Y)* in the form

$$F(x, y) = C(F_X(x), F_Y(y)), \qquad (3)$$

where F_X, F_Y are marginal distribution functions.

Let X, Y be some continuous random variables with joint distribution function $F(x,y)$ and copula C satisfying (3).

Kendall's tau for the random vector *(X, Y)* is defined (cf. [4]) by

$$\tau(X, Y) = P\{(X - \tilde{X})(Y-\tilde{Y}) > 0\} - P\{(X-\tilde{X})(Y-\tilde{Y}) < 0\}, \quad (4)$$

where $(\tilde{X}, \tilde{Y})$ is an independent copy of *(X, Y)*.

It is well know that (cf. [3])

$$\tau(X, Y) = 4 \int\int_{[0,1]^2} C(u, v)\, dC(u, v) - 1. \quad (5)$$

Some classes of bivariate copulas

The Table 1 provides a summary of basic facts (presented e.g. in [3], [7]) that are related to the families of classes Archimedean copulas that we utilize in our analyses.

Table 1: Characteristics for some Archimedean copulas

Family of copulas	Parameters	Bivariate copula C(u,v)	Kendall's τ
Gumbel	$b \geq 1$	$\exp\left\{-\left[(-\ln(u))^b + (-\ln(v))^b\right]^{1/b}\right\}$	$\frac{b-1}{b}$
Strict Clayton	$a > 0$	$\left(u^{-a} + v^{-a} - 1\right)^{-1/a}$	$\frac{a}{a+2}$
Frank	$\theta \in \Re$	$-\frac{1}{\theta}\ln\left(1 + \frac{\left(e^{-\theta u} - 1\right)\left(e^{-\theta v} - 1\right)}{\left(e^{-\theta} - 1\right)}\right)$	$1 - \frac{4}{\theta}\left(1 - D_1(\theta)\right)$

where $D_1(x) = \frac{1}{x}\int_0^x \frac{t}{e^t - 1} dt$ is the Debye function.

Fitting of copulas

In practical fitting of the data we have utilized the *maximum pseudolikelihood method* (MPLE) of parameter estimation (with initial parameters estimate received by the minimalization of the mean square distance to the empirical copula C_n presented e.g. in [4]). It requires that the copula C_θ (u, v) is absolutely continuous with density $c_\theta(u, v) = \frac{\partial^2}{\partial u\, \partial v} C_\theta(u, v)$. This method (described e.g. in [4]) involves maximizing a rank-based log-likelihood of the form

$$L(\theta) = \sum_{i=1}^{n} \ln\left(c_\theta\left(\frac{R_i}{n+1}; \frac{S_i}{n+1}\right)\right) \quad (6)$$

where n is the sample size and θ is vector of parameters in the model. Note that arguments $\frac{R_i}{n+1}, \frac{S_i}{n+1}$ equal to corresponding values of empirical marginal distributional functions of random variables X and Y..

To compare a goodness of fit of our estimated model we apply the *Akaike information criterion* in the form (see e.g. [4])

$$AIC = -2L(\boldsymbol{\theta}) + 2k \quad (7)$$

where k is the number of independent parameters in the model. Decreasing of AIC values indicate improvement of the quality of model fitting. To obtain initial values of parameters for

maximalization of the $L(\boldsymbol{\theta})$ function we apply the *mean square error method*. It is based on minimalization of the distance to the empirical copula

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} 1\left(\frac{R_i}{n+1} \le u, \frac{S_i}{n+1} \le v\right) \tag{8}$$

(with $1(\Omega)$ denoting the indicator function of set Ω) and its parametric approximation C_θ in a considered class of copulas.

Goodness of fit (GOF) test

We followed the approach of [8] and [9] for goodness of fit test measuring the size of misspecification in the form of the statistics with assymptotical distribution of the type $\chi^2_{k(k+1)/2}$ where k is the number of the estimated parameters. We use a simplified version of this statistics suggested for practical purposes in [9].

To compare goodness of fit of the models from several classes of copulas, we apply the famous Akaike criterion.

3. Results

In this section we summarized all results in tables. All calculations have been performed using the system *MATHEMATICA*, version 7. For all statistical tests that we utilized in the subsequent analyses, we applied the significance level 0.05.

For our research we used 25 real data series (exchange rates, varied macroecomic data and other financial data series).

At first, we 'fitted' these time series with SETAR model (see [10]). We based the selection of the models (optimizing the number of states and the order of the local autoregressive models) on the BIC criterion (see, e.g. [1], [5]). Recall that the residuals of these models are supposed to be independent (not only serially non-correlated). This property can be tested by the BDS test (see [2]).

Inspired by the approach of Rakonczai (2009) ([6]) we applied autocopulas to the time series of the above mentioned residuals in order to gauge how much they influence the quality of model description. If the test showed dependence in residuals, we described this dependence of time lagged residuals of SETAR models by means of copulas).

Results – The BDS test

At first we tested for linearity against the (2 regimes) SETAR type of nonlinearity. Subsequently, we selected 18 time series among those that exhibit nonlinearity (based on the above tests). Then we tested our real data series with the BDS test. Zero hypothesis is independence in residuals $\{\hat{e}_t\}$. We used significance level $\alpha = 0.05$.

The BDS test determined dependence in residuals in 9 cases (from 18) and here we used description of residual dependence with 'k-lagged auto-copula'. Correlated residuals are underlined.

Table 2. p-values for the diagnostic checking of estimate models (only for dependent time series residuals)

Time series	p-value for			
	The linearity test	**The remaining nonlinearity test**	**The correlation test**	**BDS test**
Croation Kuna	0,01613	0,01863	0,14871	0,00388
Dow Jones	0,00237	0,00874	0,15262	0,00019
Euribor 1 month	$\leq 10^{-6}$	0,01110	$\leq 10^{-6}$	$\leq 10^{-6}$
Final consumption H	0,03272	0,01135	$\leq 10^{-6}$	0,00333
GDP USA	0,00903	0,00942	0,07623	0,00884
S&P 500	0,01336	0,00021	0,69201	0,00017
UK	0,00025	0,00029	0,20150	$\leq 10^{-6}$
Unit labour cost USA	0,04171	0,00505	0,22499	$\leq 10^{-6}$
Zloty	0,04507	0,01091	$\leq 10^{-6}$	$\leq 10^{-6}$

The best copulas & Improved model SETAR

The criteria we have chosen the best copulas were values of L2 norm and information criterion AIC. The smallest value of AIC and L2 norm (from our 3 families of copulas) means the best description of residuals. For 4 time series the best copula is from Frank class, for 4 time is from Gumbel class and only in 1 case the best copula is from Clayton class.

Now we would like to improve classical SETAR model. Instead of $\{e_t\}$, (which is the strict white noise process with $E[e_t] = 0$, $D[e_t] = \sigma_e^2$ for all t = 1,..., n), we used the copulas that we have chosen as the best copulas above (for each real time series). To compare the quality of the optimal SETAR models in the categories without and with application of copulas for modeling of their residuals, we compared their standard deviations (σ_ε). The results for 9 time series are presented in Table 3. The improvement of the quality of fitting for these time series looks quite impressively.

Table 3.: Comparison of SETAR model and his improved model – description

Time series	Kendall's τ	Type of copula	Relative reduction (%) of description
Croation Kuna	0,047	Gumbel	27,5211
Dow Jones	-0,0486	Frank	17,4681
Euribor 1 month	-0,249	Frank	16,5797
Final consumption H	0,029	Clayton	0,07
GDP USA	-0,069	Frank	-21,79
S&P 500	-0,0497	Frank	26,27
UK	0,0133	Gumbel	51,11
Unit labour cost USA	0,0795	Gumbel	30,91
Zloty	0,2402	Gumbel	5,01

4. Summary

The topics of this paper was motivated by modelling of a large number of economic and financial time series from the emerging Central–European economics by SETAR model (see [1], [5]).

We have based the selection of the models (optimizing the number of states and the order of the local autoregressive models) on the BIC criterion. Recall that the residuals of these models are supposed to be independent (not only serially non-correlated). This property

can be tested e.g. by the BDS test ([2]). The BDS test has showed residual dependence (for $\alpha = 0.05$) in 9 cases (from 18) for the lag $k = 1$.

Inspired by the approach of Rakonczai [6] we have applied autocopulas to the time series of the above mentioned residuals in order to gauge how much they influence quality of description of the models Then we compared description of 'original' and 'improved' SETAR model. We have seen that when we used copulas instead of white noise in SETAR model, the descriptions were much better.

We calculated also predictions by this copula approach but we don't mention their results because it comes to the downgrade (in all cases).

In the further research we want to describe our time series with non-Archimedean copulas like Gauss, Student copulas, etc. We also want to use more complicated regime-switching models – for example STAR and MSW models.

5. References

[1] ARLT, J. – ARLTOVÁ, M. 2003. Finanční časové řady. Grada Publishing a.s., Praha. 2003

[2] BROCK, W. A. – DECHERT, W.D. – SCHEINKMAN, J.A. – LE BARON, B. 1996. A test for independence. based on the correlation dimension. In: Econometric Reviews No. 15, 1996, p. 197-235.

[3] EMBRECHTS, P. – LINDSKOG, F. – MCNAIL, A. 2001. Modeling dependence with copulas and applications to risk management. In: Rachev, S. (Ed.) Handbook of Heavy Tailed Distributions in Finance. Elsevier, Chapter 8, 2001, p. 329-384.

[4] GENEST, C. – FAVRE, A.C. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. In: Journal of Hydrologic Engineering July/August, p. 347–368

[5] FRANCES, P.H. – VAN DIJK, D. 2000. Non-linear time series models in empirical finance. Cambridge University Press, Cambridge, 2000.

[6] RAKONCZAI, P. 2009. On modeling and prediction of multivariate extremes with applications to enviromental data. In: Licentiate Theses in Mathematical Sciences, 2009:3, ISSN 1404-028X, p. 83-109.

[7] JOE, H. 1997. Models and Dependence Concepts. Chapman & Hall, London, UK, 1997.

[8] PROKHOROV, A. 2008. A goodness-of-fit test for copulas. MPRA Paper No. 9998, online at http://mpra.ub.uni-muenchen.de/9998, 2008.

[9] WHITE, H. 1982. Maximum likelihood estimation of misspecified models. In: Econometrica No. 50, 1982, p. 1-26

[10] CAMPBELL, S.D. 2002. Specification Testing and Semiparametric Estimation of Regime Switching Models: An Examination of the US Short Term Interest Rate. Department of Economics, Brown University, 2002

Adresa autora (-ov):

Anna Petričková, Mgr.
Department of Mathematics
Faculty of Civil Engineering
Slovak University of Technology Bratislava
Radlinskeho 11, 813 68 Bratislava
petrickova@math.sk

Jana Lenčuchová, Mgr.
Department of Mathematics
Faculty of Civil Engineering
Slovak University of Technology Bratislava
Radlinskeho 11, 813 68 Bratislava
lencuchova@math.sk

***Acknowledgement**: This work was supported by Slovak Research and Development Agency under contract No. LPP-0111-09.*

Vylepšený exaktný konfidenčný interval pre parameter binomického rozdelenia

The improved exact confidence interval for the binomial proportion

Ivana Pobočíková

Abstract: An interval estimation of a binomial proportion is one of the basic problem of statistics. The exact Clopper–Pearson confidence interval is a valid choice for a situation if strict coverage is required. It is known that this interval is overly conservative and too wide. In this paper we recommend the improved alternative of the Clopper–Pearson confidence interval if a strict coverage is required. The Blaker confidence interval as a strict conservative interval is less conservative and shorter than the Clopper–Pearson interval and it is a good choice in practice. We compare the performance of the confidence intervals in terms of the coverage probability, the interval length and the root mean square error.

Key words: binomial distribution, binomial proportion, confidence interval, coverage probability, interval length, root mean square error, Clopper–Pearson interval, Blaker interval

Kľúčové slová: binomické rozdelenie, konfidenčný interval, pravdepodobnosť pokrytia, dĺžka intervalu, stredná kvadratická chyba, Clopper–Pearsonov interval, Blakerov interval

1. Úvod

Medzi základné štatistické problémy patrí intervalový odhad parametra π binomického rozdelenia. Nech náhodná premenná X má binomické rozdelenie s parametrami $n \in N$ a $\pi \in \langle 0,1 \rangle$, t. j.

$$P(X = x) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, x = 0,1,...,n. \tag{1}$$

V praxi je hodnota parametra π zvyčajne neznáma a musí sa odhadnúť z výberu. Nech X znamená počet úspechov vo výbere rozsahu n. Maximálne vierohodným odhadom parametra π je $p = \frac{X}{n}$. Úlohou je nájsť $100.(1-\alpha)\%$-ný obojstranný konfidenčný interval $CI(X,n) = \langle p_L, p_U \rangle$ pre parameter π taký, že

$$P(p_L \le \pi \le p_U) \ge 1-\alpha, \tag{2}$$

kde $(1-\alpha)$ je koeficient spoľahlivosti a $\alpha \in (0,1)$. Vzhľadom na to, že binomické rozdelenie je diskrétne, je táto úloha dosť komplikovaná.

Vo väčšine štatistickej literatúry možno nájsť štandardný Waldov interval (Laplace, 1812) a exaktný Clopper-Pearsonov interval (Clopper a Pearson, 1934). Waldov interval je založený na aproximácii binomického rozdelenia normálnym rozdelením. Dolná a horná hranica $100.(1-\alpha)\%$-ného Waldovho interval sú

$$p_L = p - k_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, \qquad p_U = p + k_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}, \tag{3}$$

kde k_α je $\alpha-$kvantil normovaného normálneho rozdelenia. Waldov interval sa jednoducho počíta, je úzky, ale je o ňom známe, že má veľké problémy s pravdepodobnosťou pokrytia.

Nielen pre π blízke 0 alebo 1 je pokrytie problematické. V závislosti od n a π sa pravdepodobnosť pokrytia chová chaoticky. Mnohí autori podrobili tento interval kritike a neodporúčajú ho používať (Newcombe, 1998, Brown, Cai, Das Gupta, 2001, Pires a Amado, 2008). Clopper-Pearsonov interval je založený na exaktnom binomickom rozdelení. O tomto intervale je známe, že je príliš konzervatívny a dosť široký (Newcombe, 1998, Brown, Cai, Das Gupta, 2001, Pires a Amado, 2008). Je striktne konzervatívny, jeho pravdepodobnosť pokrytia je vždy väčšia alebo rovná ako je nominálna úroveň $(1-\alpha)$. Tento interval sa odporúča použiť v prípade, kedy treba garantovať minimálnu pravdepodobnosť pokrytia $(1-\alpha)$ (Newcombe, 1998, Pires a Amado, 2008).

V tomto príspevku ukážeme alternatívu ku Clopper-Pearsonovmu intervalu: Blakerov interval. Tento konfidenčný interval je užší, redukuje konzervativizmus Clopper-Pearsonovho intervalu a garantuje pravdepodobnosť pokrytia vždy väčšiu alebo rovnú ako je nominálna úroveň $(1-\alpha)$. Konfidenčné intervaly porovnáme v zmysle pravdepodobnosti pokrytia, konzervativizmu, dĺžky intervalu a strednej kvadratickej chyby.

2. Kritériá na vyšetrenie vlastností konfidenčných intervalov

Najskôr si zadefinujeme kritériá používané na vyšetrenie vlastností a porovnávanie konfidenčných intervalov.

Pravdepodobnosť pokrytia parametra π konfidenčným intervalom $CI(X,n)$ je pre dané $n \in N$ a $\pi \in \langle 0,1 \rangle$ je definovaná vzťahom

$$C(n,\pi) = P_\pi\left(\pi \in CI(X,n)\right) = \sum_{x=0}^{n} I(x,\pi)\binom{n}{x}\pi^x(1-\pi)^{n-x}, \tag{4}$$

kde $I(x,\pi) = \begin{cases} 1 & ak\ \pi \in CI(X,n) \\ 0 & ak\ \pi \notin CI(X,n) \end{cases}$ je indikátorová funkcia.

Konfidenčný interval je **striktne konzervatívny**, ak každé $n \in N$ a pre všetky $\pi \in \langle 0,1 \rangle$ platí $\min_\pi C(n,\pi) \geq 1-\alpha$.

Priemerná pravdepodobnosť pokrytia parametra π konfidenčným intervalom $CI(X,n)$ je definovaná vzťahom

$$AVEC(n) = \int_0^1 C(n,\pi)\,d\pi. \tag{5}$$

Konfidenčný interval je **konzervatívny v priemere**, ak $AVEC(n) \geq 1-\alpha$.

Stredná dĺžka konfidenčného intervalu je definovaná vzťahom

$$EL(n,\pi) = \sum_{x=0}^{n}\left[p_U(x,n) - p_L(x,n)\right]\binom{n}{x}\pi^x(1-\pi)^{n-x}, \tag{6}$$

kde $p_L(x,n)$, $p_U(x,n)$ sú dolné a horné hranice parciálnych konfidenčných intervalov.

Priemerná stredná dĺžka konfidenčného intervalu je definovaná vzťahom

$$AVEL(n) = \int_0^1 EL(n,\pi)\,d\pi. \tag{7}$$

Stredná kvadratická chyba je definovaná vzťahom

$$RMSE(n) = \sqrt{\int_0^1 |C(n,\pi) - (1-\alpha)|^2 \, d\pi}. \tag{8}$$

3. Alternatívy exaktných konfidenčných intervalov

Clopper–Pearsonov interval. Exaktný Clopper–Pearsonov interval (Clopper–Pearson, 1934) je založený na exaktnom binomickom rozdelení. Vznikol invertovaním obojstranného testu hypotézy $H_0 : \pi = \pi_0$ proti alternatíve $H_1 : \pi \neq \pi_0$. Pre dané $X = x$ dostaneme hranice konfidenčného intervalu $\langle p_L, p_U \rangle$ riešením rovníc

$$\begin{aligned} \sum_{k=0}^{x} \binom{n}{k} p_U^k \left(1 - p_U^k\right)^{n-k} &= \frac{\alpha}{2} \\ \sum_{k=x}^{n} \binom{n}{k} p_L^k \left(1 - p_L^k\right)^{n-k} &= \frac{\alpha}{2} \end{aligned} \tag{9}$$

Z rovníc (9) môžeme pre $0 < X < n$ vyjadriť dolnú a hornú hranicu $100.(1-\alpha)\%$-ného Clopper–Pearsonovho intervalu v tvare

$$p_L = Beta_{\frac{\alpha}{2}}(x, n-x+1), \qquad p_U = Beta_{1-\frac{\alpha}{2}}(x+1, n-x), \tag{10}$$

kde $Beta_\alpha(k_1, k_2)$ je α – kvantil $Beta$ – rozdelenia so stupňami voľnosti k_1 a k_2.

Pre $X = 0$ je $p_L = 0$ a $p_U = 1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}$. Pre $X = n$ je $p_L = \left(\frac{\alpha}{2}\right)^{\frac{1}{n}}$ a $p_U = 1$.

Clopper-Pearsonova p-hodnota sa pre hypotézu $H_0 : \pi = \pi_0$ definuje ako funkcia x a π_0 nasledovne

$$\beta_{CP}(\pi_0, x) = \min\left\{2 \min\left\{P\left(X \geq x \mid \pi_0\right), P\left(X \leq x \mid \pi_0\right)\right\}, 1\right\}. \tag{11}$$

Funkcia $\beta_{CP}(\pi, x)$ sa nazýva **konfidenčná krivka**. Konfidenčná krivka ako funkcia π pre pevné $X = x$ dáva všetky možné konfidenčné intervaly na všetkých hladinách významnosti α a pre všetky možné hodnoty π. Hranice $100.(1-\alpha)\%$-ného konfidenčného intervalu dostaneme ako priesečníky konfidenčnej krivky a priamky $y = \alpha$. Je zrejmé, že $100.(1-\alpha)\%$-ný Clopper–Pearsonov interval je množina π takých, že platí

$$\langle p_L, p_U \rangle = \{\pi, \beta_{CP}(\pi, x) \geq \alpha\}. \tag{12}$$

Blakerov interval. Tento interval (Blaker, 2000) je založený na tom, že hľadá lepšiu konfidenčnú krivku ako je $\beta_{CP}(\pi, x)$. Blakerova p-hodnota sa pre hypotézu $H_0 : \pi = \pi_0$ definuje ako funkcia x a π_0 nasledovne

Prípad 1. Ak $P\left(X \geq x \mid \pi_0\right) < P\left(X \leq x \mid \pi_0\right)$, potom

$$\beta_B(\pi_0, x) = P\left(X \geq x \mid \pi_0\right) + P\left(X \leq x^* \mid \pi_0\right), \tag{13}$$

kde x^* je najväčšie $u \in N$ také, že $P\left(X \leq u \mid \pi_0\right) \leq P\left(X \geq u \mid \pi_0\right)$.

Prípad 2. Ak $P\left(X \geq x \mid \pi_0\right) > P\left(X \leq x \mid \pi_0\right)$, potom

$$\beta_B(\pi_0, x) = P\left(X \leq x \mid \pi_0\right) + P\left(X \geq x^{**} \mid \pi_0\right), \tag{14}$$

kde x^{**} je najmenšie $u \in N$ také, že $P(X \geq u \mid \pi_0) \leq P(X \leq u \mid \pi_0)$.

Prípad 3. Ak $P(X \geq x \mid \pi_0) = P(X \leq x \mid \pi_0)$, potom

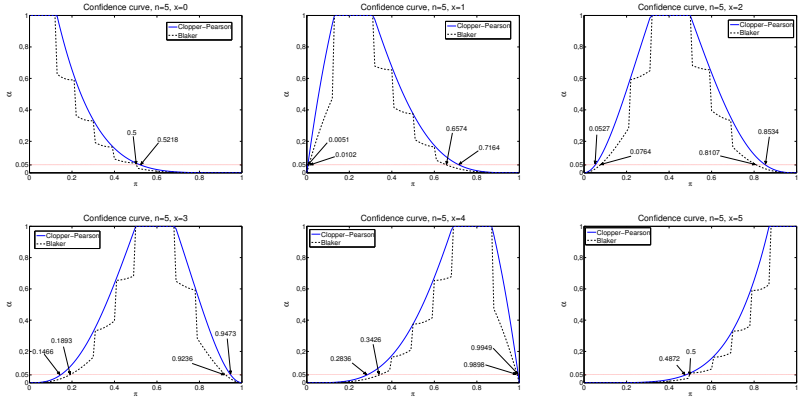
$$\beta_B(\pi_0, x) = 1. \tag{15}$$

100.$(1-\alpha)$% -ný Blakerov interval je teda množina π takých, že platí

$$\langle p_L, p_U \rangle = \{\pi, \beta_B(\pi, x) \geq \alpha\}. \tag{16}$$

V [2] možno nájsť program v jazyku S-Plus na výpočet tohto intervalu.

Na obrázku 1 sú znázornené konfidenčné krivky pre Clopper-Pearsonov a Blakerov interval pre $n = 5$ a $X = 0,1,2,3,4,5$. Priesečníkmi konfidenčnej krivky a priamky $y = 0,05$ sú dolné a horné hranice 95% konfidenčných intervalov. Z obrázku 1 je vidieť, že Blakerov interval je užší a je vždy podmnožinou Clopper-Pearsonovho intervalu.



Obrázok 8: Konfidenčné krivky pre $n=5$ a $X=0,1,2,3,4,5$

4. Porovnanie konfidenčných intervalov

Na ohodnotenie kvality konfidenčného intervalu sú kľúčové nasledujúce kritériá- pravdepodobnosť pokrytia, konzervatívnosť a dĺžka intervalu. Prednosť sa dáva konfidenčnému intervalu, ktorý má pravdepodobnosť pokrytia blízko nominálnej úrovne, je užší, má menšiu priemernú pravdepodobnosť pokrytia a menšie kolísanie pravdepodobnosti pokrytia okolo nominálnej úrovne.

Pravdepodobnosti pokrytia 95%-ných konfidenčných intervalov sa počítali v 2001 hodnotách π z intervalu $\langle 0,1 \rangle$ daných $\pi = 0,0005i, i = 0,1,...,2000$, pre $n = 1$ až 500. Všetky výpočty sa realizovali pomocou Matlabu. Pravdepodobnosti pokrytia 95%-ných konfidenčných intervalov pre $n = 50$ sú znázornené na obrázku 2. Minimálna pravdepodobnosť pokrytia 95%-ného Clopper-Pearsonovho intervalu pre $n = 1,2...,500$ je znázornená na obrázku 3. Clopper-Pearsonov i Blakerov interval sú striktne konzervatívne, garantujú minimálnu pravdepodobnosť pokrytia $(1-\alpha)$. Clopper-Pearsonov interval je však príliš konzervatívny. Pravdepodobnosť pokrytia sa zhora ťažko približuje k nominálnej

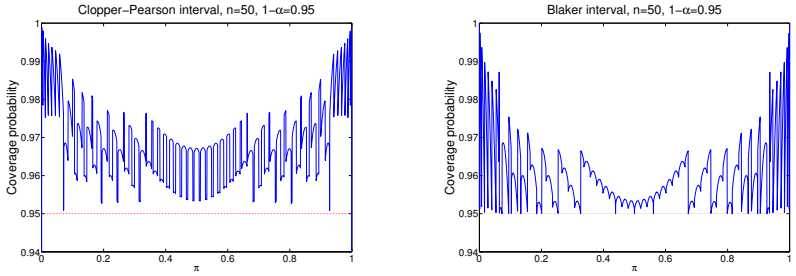
úrovni. Pre malé n je pravdepodobnosť pokrytia medzi $(1-\alpha/2)$ a 1 (Blaker, 2000). Oba intervaly sú konzervatívne v priemere (obrázok 4). Clopper-Pearsonov interval má priemernú pravdepodobnosť pokrytia hodne nad nominálnou úrovňou. Z porovnania intervalov vidieť, že Blakerov interval je menej konzervatívny a je oveľa menej konzervatívny v priemere než Clopper-Pearsonov interval. Blakerov interval je užší, má menšiu priemernú strednú dĺžku (obrázok 5) a menšiu strednú kvadratickú chybu (obrázok 6).

Tabuľka 1: Porovnanie 95% konfidenčných intervalov v zmysle minimálnej pravdepodobnosti pokrytia (MCP), priemernej pravdepodobnosti pokrytia (AVEC), strednej kvadratickej chyby (RMSE) a priemernej dĺžky intervalu (AVEL) pre $n = 10,30,50,100,500$

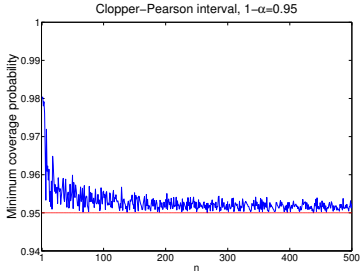
	Metóda	MCP	AVEC	RMSE	AVEL
	Clopper-Pearson	0,9610	0,9837	0,0349	0,5085
n=10	Blaker	0,9500	0,9733	0,0260	0,4760
	Clopper-Pearson	0,9506	0,9734	0,0256	0,2990
n=30	Blaker	0,9500	0,9631	0,0162	0,2833
	Clopper-Pearson	0,9509	0,9692	0,0215	0,2306
n=50	Blaker	0,9500	0,9603	0,0132	0,2206
	Clopper-Pearson	0,9504	0,9646	0,0169	0,1614
n=100	Blaker	0,9500	0,9578	0,0102	0,1566
	Clopper-Pearson	0,9503	0,9573	0,0091	0,0706
n=500	Blaker	0,9500	0,9536	0,0054	0,0695

5. Záver

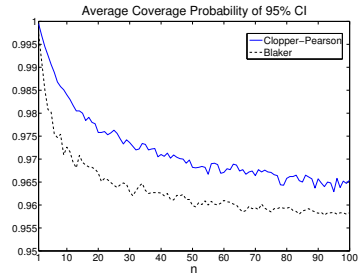
Striktne konzervatívne intervaly sa odporúčajú použiť v prípadoch, kedy treba garantovať minimálnu pravdepodobnosť pokrytia $(1-\alpha)$. Prednosť sa dáva intervalom, ktorých pravdepodobnosť pokrytia je mierne nad nominálnou úrovňou a sú užšie. Z porovnania konfidenčných intervalov je zrejmé, že Blakerov interval je menej konzervatívny ako Clopper-Pearsonov interval. Má lepšiu priemernú pravdepodobnosť pokrytia, je menej konzervatívny v priemere, má lepšiu strednú kvadratickú chybu a je užší. Je teda veľmi dobrou alternatívou Clopper-Pearsonovho intervalu na použitie v praxi.



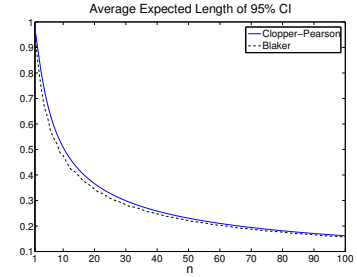
Obrázok 2: Pravdepodobnosti pokrytia 95% konfidenčných intervalov pre n=50



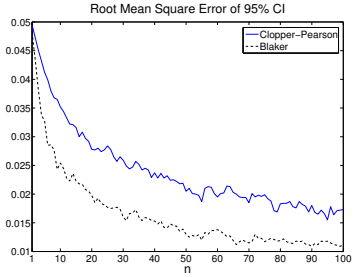
Obrázok 3: Minimálna pravepodobnosť pokrytia 95% Clopper-Pearsonovho intervalu pre $n = 1, 2..., 500$



Obrázok 4: Priemerná pravdepodobnosť pokrytia 95% konfidenčných intervalov pre $n = 1, 2..., 100$



Obrázok 5: Priemerná stredná dĺžka 95% konfidenčných intervalov pre $n = 1, 2..., 100$



Obrázok 6: Stredná kvadratická chyba 95% konfidenčných intervalov pre $n = 1, 2..., 100$

6. Literatúra

[1] BLAKER, H., 2000. Confidence curves and improved exact confidence intervals for discrete distributions. In: The Canadian Journal of Statistics 28, 2000, s. 783-798.

[2] BROWN, D. L. – CAI, T. T. – DASGUPTA, A. 2001. Interval estimation for a binomial proportion. In: Statistical Science 16, 2001, s. 101-133.

[3] CLOPPER, C. – PEARSON, S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. IN: Biometrika 26, 1934, s. 404-413.

[4] LAPLACE, P. S. 1812. Theorie analytique des probabilities. Paris, France, Courier, 1812.

[5] NEWCOMBE, R. G. 1998. Two-sided confidence intervals for the single proportion; comparison of several methods. In: Statistics in Medicine 17, 1998, s. 857-872.

[6] PIRES, M. A. – AMADO, C. 2008. Interval estimators for a binomial proportion: comparison of twenty methods. In: REVSTAT-Statistical Journal, Vol. 6, No. 2, 2008, s. 165-197.

[7] POBOČÍKOVÁ, I. 2010. Better confidence intervals for a binomial proportions. In: Communications:Scientific Letters of the University of Žilina, Vol. 12, No. 3, s.31-37.

[8] WILSON, E. B. 1937. Probable inference, the law of succession, and statistical inference. In: Journal of the American Statistical Association 22, 1927, s. 209-212.

Adresa autora:

Ivana Pobočíková, Mgr., E-mail: ivana.pobocikova@fstroj.uniza.sk
Katedra aplikovanej matematiky SjF, Žilinská univerzita, Univerzitná 1, 010 26 Žilina

Simulácia pravdepodobnosti v kartovej hre – Poker (Monte Carlo simulácia)
Simulation of probability in the Poker (Monte Carlo simulation)

Renáta Prokeinová

Abstract: Gambling is very popular belong to very favourite activity, gambling is exciting because you never know what`s going to happen. Monte Carlo simulation is a powerful tool that can be used to predict gambling probabilities. Actual method represents very strong tool, which will be used to predict of probability coming the Winn in the gambling. Essentially, we predict probability by playing out our gambling event situation multiple times.

The main goal of the paper presents application of method Monte Carlo at the prediction of coming situation "three of kind" in the poker. All calculations are executed in Excel 2007.

Key words: gambling, the poker, method Monte Carlo

Kľúčové slová: stávkovanie, poker, metóda Monte Carlo

JEL classification: C70, C78

1. Úvod

Pre lepšiu ilustráciu hry je vhodné uviesť základné pravidlá spomínanej hry. Princíp hry spočíva v tom, že každá kombinácia kariet je hierarchicky zatriedená, aby bolo možné určiť víťaza (napr. tri karty s rovnakým symbolom majú vyššiu hodnotu než dve karty s rovnakým symbolom).

Poker sa hrá s balíčkom francúzskych kariet (52) (v niektorých verziách pokru sa používajú aj žolíky). Hodnota kariet podľa veľkosti: eso, kráľ, dáma, dolník, 10, 9, 8, 7, 6, 5, 4, 3, 2 a eso. Sú 4 farby: piky, srdcia, kára a trefy. V niektorých verziách pokru sú dvojky divoké karty, ktoré môžu zastúpiť ľubovoľnú kartu.

Základné pojmy a pravidlá Poker

V pokri je často používaná metóda blufovania na zmetenie súpera. Call alebo check iba na dorovnanie pričom ste si istý, že máte najlepšiu kombináciu, avšak nechcete protivníka odstrašiť. Koniec stávkového kola nastáva vždy keď všetci hráči foldnú, checknú alebo dajú call. Ak išlo o posledné stávkové kolo, ukážu sa karty, a ten kto má najlepšiu kombináciu vyhráva.

Výherné kombinácie

Najdôležitejšou vecou je poradie výherných kombinácií. Sú zoradené od najvyššej výhernej kombinácie po najnižšiu.

Royal Flush - Ak máte postupku v jednej farbe od 10 po A, čiže 10-J-Q-K-A, stavte všetko čo máte. Je to totiž neporaziteľná kombinácia.

Straight Flush - Akákoľvek iná farebná postupka, napríklad 7-8-9-10-J v srdciach.

4 of a Kind - U nás tiež volaný "Poker". 4 karty jednej hodnoty. Napríklad: 9-9-9-9-A.

Full House - 3 karty jednej hodnoty a ďalšie 2 karty inej rovnakej hodnoty. Napríklad: 7-7-7-5-5 alebo 10-10-10-A-A.

Flush - Farba - 5 kariet z jednej farby. Ak sa stane, že majú dvaja hráči flush, vyhráva ten kto má vyššiu kartu z tej farby.

Straight - Je to postupka zložená z 5 kariet, pričom nie sú z jednej farby. Napríklad: 8-9-10-J-Q. Voláme ju tiež špinavá postupka.
3 of a Kind - 3 karty rovnakej hodnoty. Napríklad K-K-K-7-9 alebo Q-Q-Q-6-A
Two Pair - 2 páry. Sú to dve dvojice, ktoré majú rovnaké hodnoty. Napríklad 5-5-6-6-A alebo J-J-K-K-10
Pair - Jedna dvojica rovnakej hodnoty. Napríklad: 10-10-7-8-K
High Card - Ak nemá nik ani jednu z uvedených kombinácií, vyhráva hráč s vyššou kartou.
Kicker - Pri výherných kombináciách treba ešte spomenúť tzv. Kickera. Je to karta, ktorá rozhoduje pri hre, keď majú obaja hráči napríklad rovnaké 2 páre, trojicu, a pod. Príklad: ak prvý hráč má 8-8-7-7-J a druhý hráč 8-8-7-7-K vyhráva hráč druhý, ktorý má vyššieho kickera. V tomto prípade K.

Existuje množstvo pravdepodobností použitých na simuláciu odhadu pravdepodobnosti nastatia javu „three of kind“, ktorý požaduje aby sme mali tri karty jedného typu a žiaden pár. V Exceli nasimulujeme ťahanie piatich kariet použitím metódy Monte Carlo.

2. Metodika

Metódu počítačovej simulácie, ktorá je svojím charakterom blízka fyzikálnym experimentom a zároveň patrí medzi metódy numerickej matematiky, nazývame Monte Carlo. Metóda dostala názov podľa Monte Carla, známeho svojimi kasínami a najmä ruletou. Termín prvýkrát použili fyzici pracujúci na zostrojení atómovej bomby okolo roku 1940.

Metóda Monte Carlo nám pomáha riešiť pravdepodobnostné i deterministické úlohy prostredníctvom veľakrát opakovaných náhodných pokusov na vstupnej vzorke dát. Riešenie vychádza z predpokladu, že zostrojíme pravdepodobnostnú úlohu, ktorá má rovnaké riešenie s pôvodnou úlohou. Dosiahnuté riešenie má pravdepodobnostný charakter.

Výnimočnosť tejto metódy spočíva v tom, že výpočtové operácie sa realizujú so pseudonáhodnými číslami, čo má za následok nedeterminovanosť výpočtového procesu z hľadiska výsledku. Ide o špecifickú techniku výberu vzoriek. Všeobecne vzorkovanie chápeme ako proces, pomocou ktorého vstupné parametre pre ľubovoľný model možno vybrať zo vstupných hodnôt distribučnej funkcie.

Túto metódu vo všeobecnosti definujeme ako numerickú metódu, ktorá pomáha pri vytváraní a analyzovaní matematických modelov rôznych procesov, ktorých priebeh je ovplyvňovaný náhodnými faktormi. V závislosti od objektu výskumu, ktorým sú procesy, v ktorých správaní sa náhodné deje vo veľkej miere prejavujú, je táto metóda pomerne rozšírená. Tento postup však nie je obmedzený len na riešenie úloh, ktoré majú náhodný charakter. Dôkazom tohto tvrdenia je množstvo deterministických úloh, ktoré v podstate nespájame so žiadnymi náhodnosťami (napríklad riešenie sústav lineárnych rovníc, riešenie známej Laplaceovej rovnice atď.) a pre ktoré možno vytvoriť pravdepodobnostné modely na ich riešenie.

Metóda Monte Carlo sa okrem matematických problémov používa napríklad na riešenie problémov v ekonómii, najmä vo finančníctve a poisťovníctve. Vo financiách bola po prvýkrát využitá v roku 1977 pri riešení problémov s oceňovaním finančných aktív. V súčasnosti sa používa na simuláciu rôznych zdrojov neistoty a rizika, pri ohodnotení portfólia, oceňovaní opcií. Konkrétne môžeme simuláciu použiť pri odhade priemerného výnosu a rizikovosti nových produktov, vieme pomocou nej predpovedať čistý zisk pre spoločnosť, výšku štrukturálnych nákladov a nákladov na nákup a takisto určiť citlivosť na riziká, ako je zmena úrokových sadzieb, kolísanie výmenného kurzu.

Vo všeobecnosti je metóda Monte Carlo založená na hľadaní hodnoty veličiny x, ktorú určitým spôsobom charakterizuje náhodný proces. Tento proces simulujeme na počítači a

získame tak realizácie náhodnej veličiny ˇx. Na základe týchto realizácií veličiny ˇx odhadujeme s určitou presnosťou hodnotu hľadanej veličiny x.

3. Metóda počítačovej simulácie v MS Exceli.

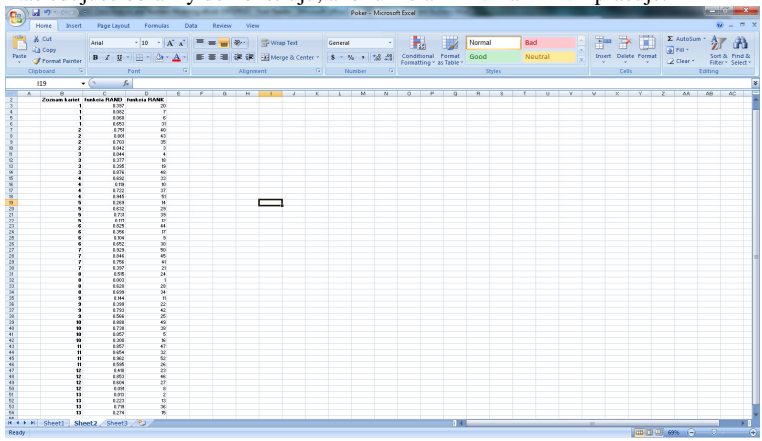
Aká je pravdepodobnosť získania „three of kind“?

Na simuláciu modelov rôznych náhodných procesov je najlepšie použiť široko dostupný nástroj od Microsoftu – Excel. Na jednoduchých príkladoch, ktoré riešia finančné otázky, si ukážeme, ako si MS Excel dokáže poradiť s náhodnými číslami a pravdepodobnosťou.

Začneme tým, že si do Excelu nahodíme pod seba všetky hodnoty kariet v balíčku, teda štyri jednotky, štyri dvojky až po štyri esá, ktoré sú prezentované číslom 13. Do ďalšieho stĺpca sa nastavíme vedľa prvej hodnoty prvej karty v balíčku. Vyvoláme si funkciu RAND (), použitím ktorej získame náhodné číslo s normálnym rozdelením. Spomínané náhodné číslo môže nadobúdať hodnotu z intervalu od 0 do 1.

Následne je potrebné identifikovať poradie vytiahnutia z balíčka pre každú kartu. Poradie získame použitím funkcie RANK(). Syntax funkcie RANK je nasledovná. RANK (číslo, rozsah, 0 alebo 1). Budeme sa hlbšie venovať poslednému parametru. Ak posledný parameter zadáme hodnotu 1, funkcia poradie pridá jednotlivým číslam poradie najmenšie čísla blízke hodnote 1. Ak zadáme do posledného parametra 0, získame najvyššie čísla blízke hodnote 1.

Nasledujúce obrázky demonštrujú, ako funkcia RAND a RANK pracuje.

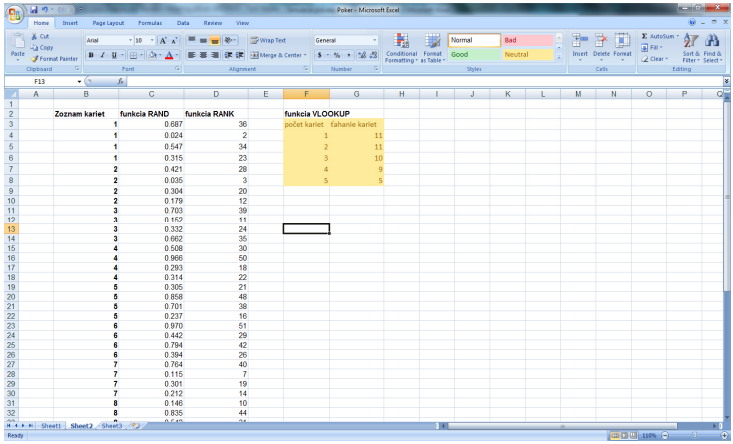


Obrázok 1 Simulácia diskrétnej náhodnej premennej a demonštrácia funkcie RANK a RAND

Zdroj: Vlastné výpočty

V danom momente máme pripravenú simuláciu poradia ťahania kariet z balíčka. Ďalší krok pozostáva zo simulácie výberu piatich kariet, pričom využijeme tabuľku nasimulovaných poradí, získanú cez funkciu RANK().

Pre vhodnú simuláciu výberu piatich kariet použijeme fukciu VLOOKUP(). Syntax funkcie VLOOKUP (hodnota, pre ktoré chceme zistiť poradie; lookup - tabuľka vopred vypočítaných poradí; číslo stĺpca kde sa nachádza poradie; FALSE). Vzorec vyberie päť kariet ktoré korešpondujú s piatimi najmenšími náhodnými číslami.



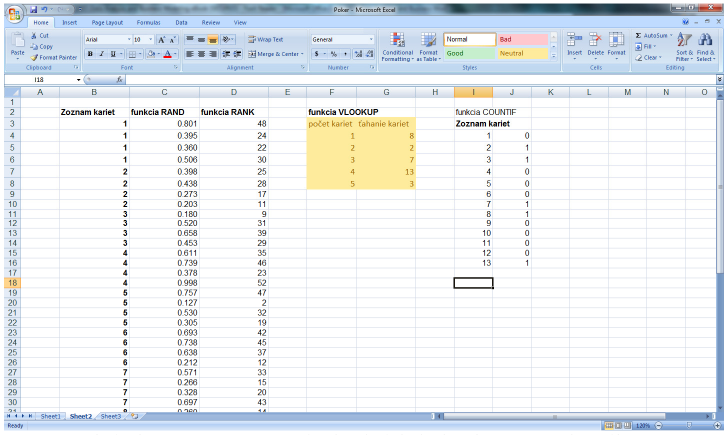
	Zoznam kariet	funkcia RAND	funkcia RANK		funkcia VLOOKUP	
					počet kariet	ťahanie kariet
	1	0.687	36		1	11
	1	0.024	2		2	11
	1	0.547	34		3	10
	1	0.315	23		4	9
	2	0.421	28		5	5
	2	0.035	3			
	2	0.304	20			
	2	0.179	12			
	3	0.703	39			
	3	0.159	11			
	3	0.332	24			
	3	0.662	35			
	4	0.508	30			
	4	0.966	50			
	4	0.293	18			
	4	0.314	22			
	5	0.305	21			
	5	0.858	48			
	5	0.701	38			
	5	0.237	16			
	6	0.970	51			
	6	0.442	29			
	6	0.794	42			
	6	0.394	26			
	7	0.764	40			
	7	0.115	7			
	7	0.301	19			
	7	0.212	14			
	8	0.146	10			
	8	0.835	44			

Obrázok 2 Výber piatich kariet na základe pravdepodobného poradia cez funkciu VLOOKUP

Zdroj: Vlastné výpočty

Keď už máme zaznamenané poradie kariet, ktoré budú ťahané (cez funkciu VLOOKUP), môžeme pristúpiť k zisteniu početností pre každú skupinu kariet nachádzajúcich sa v balíčku. Identifikácia početností je úzko previazaná s predchádzajúcou tabuľkou, v ktorej sme simulovali aké druhy kariet môže hráč vytiahnuť, ak ťahá iba päť kariet. V závislosti od toho aké karty vytiahne, funkcia COUNTIF nám ich napočíta a roztriedi do skupín, presne tých istých skupín, ktoré sme si na začiatku navolili pod hlavičkou "Zoznam kariet".

Funkcia COUNTIF() má nasledovnú syntax COUNTIF(DRAWN - stĺpec ťahaných kariet; hodnota danej karty zo zoznamu kariet - stanovená trieda diskrétneho znaku).

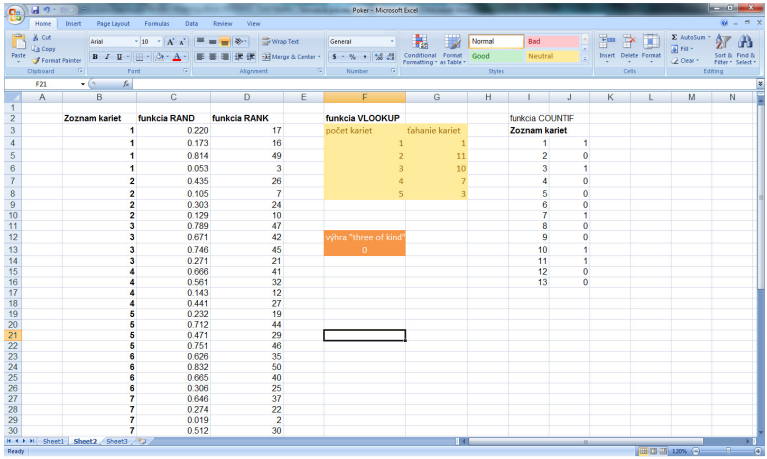


	Zoznam kariet	funkcia RAND	funkcia RANK		funkcia VLOOKUP			funkcia COUNTIF	
					počet kariet	ťahanie kariet		Zoznam kariet	
	1	0.801	48		1	8		1	0
	1	0.395	24		2	2		2	1
	1	0.360	22		3	7		3	1
	1	0.506	30		4	13		4	0
	2	0.398	25		5	3		5	0
	2	0.438	28					6	0
	2	0.273	17					7	1
	2	0.203	11					8	1
	3	0.180	9					9	0
	3	0.520	31					10	0
	3	0.658	39					11	0
	3	0.453	29					12	0
	4	0.611	35					13	1
	4	0.739	46						
	4	0.378	23						
	4	0.998	52						
	5	0.757	47						
	5	0.127	2						
	5	0.530	32						
	5	0.305	19						
	6	0.693	42						
	6	0.738	45						
	6	0.638	37						
	6	0.212	12						
	7	0.571	33						
	7	0.266	15						
	7	0.328	20						
	7	0.697	43						

Obrázok 3 Zistenie početností jednotlivých druhov kariet pri jedno výbere

Zdroj: Vlastné výpočty

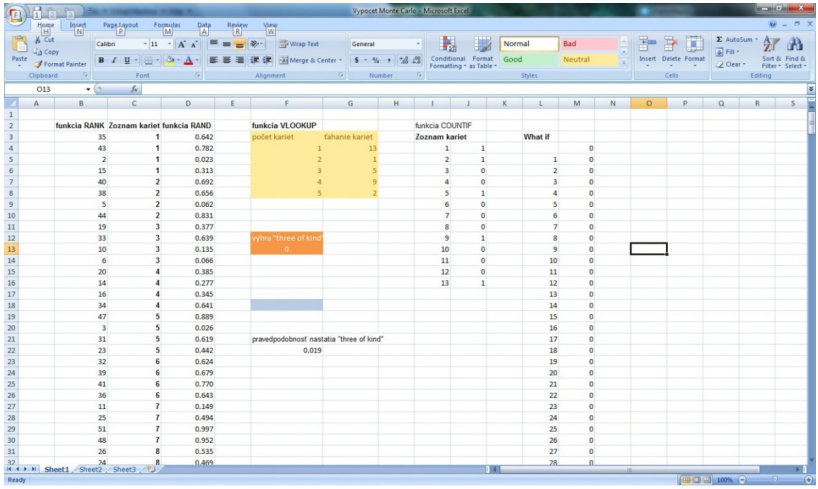
Z poslednej tabuľky môžeme identifikovať, či daný ťah piatich kariet obsahuje tri karty jedného druhu a dve rôzne karty. Môžeme využiť funkciu IF() s nasledovným syntaxom. =IF(AND(MAX(J4:J16)=3,COUNTIF(J4:J16,2)=0),1,0).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		Zoznam kariet	funkcia RAND	funkcia RANK		funkcia VLOOKUP			funkcia COUNTIF					
3		1	0.220	17		počet kariet	ťahanie kariet		Zoznam kariet					
4		1	0.173	16		1	1		1	1				
5		1	0.814	49		2	11		2	0				
6		1	0.053	3		3	10		3	1				
7		2	0.435	26		4	7		4	0				
8		2	0.105	7		5	3		5	0				
9		2	0.303	24					6	0				
10		2	0.129	10					7	1				
11		3	0.789	47					8	0				
12		3	0.671	42		vyhra "three of kind"			9	0				
13		3	0.746	45		0			10	1				
14		3	0.271	21					11	1				
15		4	0.666	41					12	0				
16		4	0.561	32					13	0				
17		4	0.143	12										
18		4	0.441	27										
19		5	0.232	19										
20		5	0.712	44										
21		5	0.471	29										
22		5	0.751	46										
23		6	0.626	35										
24		6	0.832	50										
25		6	0.665	40										
26		6	0.306	25										
27		7	0.646	37										
28		7	0.274	22										
29		7	0.019	2										
30		7	0.512	30										

Obrázok 4 Podmienka IF, ktorá identifikuje výber kariet „three of kind“
Zdroj: Vlastné výpočty

V tomto prípade nenastala žiadna zmožností ťahu kariet, kde by vznikla situácia tri karty jedného druhu a dve rôzne karty. Vzorec v oranžovom políčku nadobudne hodnotu 1 iba a len v tom prípade, ak nastane situácia "three of kind".



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1																			
2		funkcia RANK	Zoznam kariet	funkcia RAND		funkcia VLOOKUP			funkcia COUNTIF										
3		35	1	0.642		počet kariet	ťahanie kariet		Zoznam kariet			What if							
4		43	1	0.782		1	13		1	1			0						
5		2	1	0.023		2	1		2	1		1	0						
6		15	1	0.313		3	5		3	0		2	0						
7		40	2	0.692		4	9		4	0		3	0						
8		38	2	0.656		5	2		5	1		4	0						
9		5	2	0.062					6	0		5	0						
10		44	2	0.831					7	0		6	0						
11		19	3	0.377					8	0		7	0						
12		33	3	0.639		vyhra "three of kind			9	1		8	0						
13		10	3	0.135		0			10	0		9	0						
14		6	3	0.066					11	0		10	0						
15		20	4	0.385					12	0		11	0						
16		14	4	0.277					13	1		12	0						
17		16	4	0.345								13	0						
18		34	4	0.641								14	0						
19		47	5	0.889								15	0						
20		3	5	0.026								16	0						
21		31	5	0.619		pravdepodobnost nastatia "three of kind"						17	0						
22		23	5	0.442		0,019						18	0						
23		32	6	0.624								19	0						
24		39	6	0.679								20	0						
25		41	6	0.770								21	0						
26		36	6	0.643								22	0						
27		11	7	0.149								23	0						
28		25	7	0.494								24	0						
29		51	7	0.997								25	0						
30		48	7	0.952								26	0						
31		26	8	0.535								27	0						
32		24	8	0.469								28	0						

Obrázok 5 Simulácia pravdepodobnosti nastatia javu „three of kind“-4000 iterácii
Zdroj: Vlastné výpočty

V tomto momente máme simuláciu kompletne nastavenú a môžeme simulovať 4000 pokerových hier. Pre simuláciu rozsiahleho počtu opakovaní využijeme jeden z analytických nástrojov What if, ktorý nám vygeneruje jednostupňovú tabuľku. Bunku M4 prepojíme vzorcom =F13. Následne označíme do bloku oblasť buniek L4:M4004, môžeme spustiť funkciu What if a vyberieme si možnosť Data Table (tabuľka údajov). V dialógovom okne máme dve možnosti, vytvoriť tabuľku v riadkoch alebo v stĺpcoch. V našom prípade sa nastavíme možnosť vytvoriť stĺpcovú tabuľku. Vyberieme prázdnu bunku ako vstupnú bunku stĺpca. Po odkliknutí funkcie sa vygenerujú možnosti nastatia skúmanej situácie pre opakovaný počet 4000 krát.

Poslednou funkciou ukončíme simuláciu. Vytvoríme priemernú hodnotu zo všetkých iterácii, vygenerovaných funkciou What if. V prvej možnosti, často nastáva nulová pravdepodobnosť nastatia javu. Po následnom klikaní funkčnej klávesy F9 sa situácia mení a generujú sa nám pravdepodobnosti, ktoré sú vzorcom prepojené s celým modelom simulácie.

4. Záver

Na základe ukážkového príkladu s použitím metódy Monte Carlo boli demonštrované možnosti jej využitia. Metóda Monte Carlo sa právom považuje za univerzálnu metódu riešenia matematických problémov, ako aj úloh z oblasti ekonómie. Finančníci dokážu napríklad stanoviť optimálnu investičnú stratégiu pre dôchodok svojich klientov.

Okrem funkcií programu MS Excel môžeme nasimulovať rôzne procesy aj pomocou nasledujúcich doplnkov k Excelu: Risk Solver, @Risk, Risk Analyzer, MC Add-In, MonteCarlito.

5. Literatúra

[1]FABIAN, F. – KLUIBER, Z. 1998. Metoda Monte Carlo-její vznik a uplatnení. Metoda Monte Carlo a možnosti jejího uplatnění Journal of Econometrics.1998.ISBN 80-7175-058-1

[2]WINSTON, W.L. 2007. Fun and Games: Simulating Gambling and Sporting Event Probabilities. Microsoft Office Excel 2007. ISBN 2007924648, s.501-505

Adresa autora (-ov):

Renáta Prokeinová, Ing., PhD.
Fakulta ekonomiky a manažmentu
Slovenská poľnohospodárska univerzita
Tr. A. Hlinku 2, 949 01 Nitra
renata.prokeinova@fem.uniag.sk

Porovnanie odhadov variančných parametrov v špeciálnom modeli rastových kriviek

The comparison of two estimators of variance parameters in a special growth curve model

Rastislav Rusnačko

Abstract: In this text, we compare the estimators of variance parameters in the growth curve model with uniform correlation structure derived by Zezula on the one side and Ye and Wang on the other side. We investigate properties of difference of mean square error of the two estimators in different situations.

Key words: Growth Curve Model, uniform correlation structure, generalized uniform correlation structure

Kľúčové slová: Model rastových kriviek, rovnomerná korelačná štruktúra, zovšeobecnená rovnomerná korelačná štruktúra

JEL classification: C13, C29, C39

1. Úvod

Štandardný model rastových kriviek predstavili Potthoff a Roy už v roku 1964. Tento model sa zvykne označovať aj ako zovšeobecnený model viacrozmernej analýzy variancie, zovšeobecnený lineárny model alebo aj Potthoffov a Royov model. Neskôr sa objavili rôzne modifikácie tohto modelu, o.i. model s rovnomernou korelačnou štruktúrou. Odhady variančných parametrov v tomto modeli odvodil Žežula a neskôr aj Ye a Wang. V tomto texte porovnáme tieto odhady na základe ich stredných kvadratických chýb.

2. Model rastových kriviek

Štandardný model rastových kriviek je tvaru

$$Y = XBZ + \varepsilon, \quad E\varepsilon = 0, \quad \mathrm{var}(\mathrm{vec}\,\varepsilon) = \Sigma \otimes I, \tag{1}$$

kde $Y_{n\times p}$ je matica pozorovaní, $X_{n\times m}$ je matica analýzy variancie, $Z_{r\times p}$ je matica regresných konštánt, $B_{m\times r}$ je matica neznámych parametrov, $\varepsilon_{n\times p}$ je matica náhodných chýb a I je jednotková matica. Matica Σ opisuje závislosť riadkov matice Y a vec operátor poskladá stĺpce matice pod seba a vytvorí tak z matice stĺpcový vektor.

Pretože variančná matica Σ obsahuje veľa neznámych parametrov, je niekedy užitočné ich počet redukovať uvažovaním jednoduchej štruktúry. Jednou z nich je model s rovnomernou korelačnou štruktúrou:

$$\Sigma = \sigma^2[(1-\rho)I + \rho\mathbf{1}\mathbf{1}']. \tag{2}$$

3. Rovnomerná korelačná štruktúra

Keďže $E(S) = \Sigma$, tak $E[\mathrm{Tr}(S)] = p\sigma^2$ a $E(\mathbf{1}S\mathbf{1}') = p\sigma^2[1 + (p-1)\rho]$. Rovnice

$$\mathrm{Tr}(S) - p\hat{\sigma}^2 = 0 \quad \text{a} \quad \mathbf{1}'S\mathbf{1} - \mathrm{Tr}(S)[1 + (p-1)\hat{\rho}] = 0 \tag{3}$$

sú potom nevychýlené odhadovacie rovnice parametrov σ^2 a ρ. Z toho hneď vyplýva, že prirodzené odhady sú

$$\hat{\sigma}^2 = \frac{\mathrm{Tr}(S)}{p} \quad a \quad \hat{\rho} = \frac{1}{p-1}\left(\frac{\mathbf{1}'S\mathbf{1}}{\mathrm{Tr}(S)} - 1\right). \tag{4}$$

Tieto odhady odvodil Žežula a neskôr v inej (ekvivalentnej) podobe Ye a Wang.

4. Zovšeobecnená rovnomerná korelačná štruktúra

Niekedy sa uvažuje korelačná štruktúra v tvare

$$\Sigma = \theta_1 G + \theta_2 w w', \tag{5}$$

pričom G je známa symetrická matica a *w* je známy vektor. Keďže Σ musí byť pozitívne semidefinitná, tak rozlišujeme dva prípady. Keď $\theta_1 \geq 0$ a $\theta_2 \leq 0$, tak

$$\theta_1 G - |\theta_2| w w' \geq 0 \iff G \geq 0, \quad w \in R(G), \quad |\theta_2| w' G^{+} w \leq \theta_1. \tag{6}$$

V prípade, keď $\theta_1 \geq 0$ a $\theta_2 \geq 0$, tak $\theta_1 G + \theta_2 w w' \geq 0$ pre každý vektor *w*. Podobne ako v modeli rovnomernej korelačnej štruktúry platí

$$E[Tr(S)] = \theta_1 Tr(G) + \theta_2 w' w \quad \text{a} \quad E(\mathbf{1}'S\mathbf{1}) = \theta_1 \mathbf{1}'G\mathbf{1} + \theta_2 (\mathbf{1}'w)^2. \tag{7}$$

Na základe toho, odhady odvodené Žežulom sú v tvare

$$\hat{\theta}_1^Z = \frac{(\mathbf{1}'w)^2 Tr(S) - \mathbf{1}'S\mathbf{1} w'w}{(\mathbf{1}'w)^2 Tr(G) - \mathbf{1}'G\mathbf{1} w'w} \quad \text{a} \quad \hat{\theta}_2^Z = \frac{\mathbf{1}'S\mathbf{1} Tr(G) - \mathbf{1}'G\mathbf{1} Tr(S)}{(\mathbf{1}'w)^2 Tr(G) - \mathbf{1}'G\mathbf{1} w'w}. \tag{8}$$

Na druhej strane, Ye a Wang odvodili odhady v tvare

$$\hat{\theta}_1^{YW} = \frac{w'G^{-1} w Tr(G^{-1}S) - w'G^{-1} S G^{-1} w}{(p-1)(w'G^{-1}w)} \quad \text{a} \quad \hat{\theta}_2^{YW} = \frac{p w'G^{-1} S G^{-1} w - w'G^{-1} w Tr(G^{-1}S)}{(p-1)(w'G^{-1}w)^2}. \tag{9}$$

Keďže všetky tieto odhady sú nevychýlené, teda platí

$$E(\hat{\theta}_1^Z) = \theta_1, \qquad E(\hat{\theta}_2^Z) = \theta_2, \qquad E(\hat{\theta}_1^{YW}) = \theta_1 \quad \text{a} \quad E(\hat{\theta}_2^{YW}) = \theta_2, \tag{10}$$

tak ich stredné štvorcové chyby sú rovnaké ako ich variancie. Výpočtom dostávame, že pre ne platia vzťahy

$$MSE\,\hat{\theta}_1^Z = Var\,\hat{\theta}_1^Z = \frac{2}{n - r(X)} \cdot \frac{(\mathbf{1}'w)^4 Tr(\Sigma^2) - 2(\mathbf{1}'w)^2 w'w \mathbf{1}'\boldsymbol{\Sigma}^2\mathbf{1} + (w'w)^2 (\mathbf{1}'\Sigma\mathbf{1})^2}{[(\mathbf{1}'w)^2 Tr(G) - \mathbf{1}'G\mathbf{1} w'w]^2}, \tag{11}$$

$$MSE\,\hat{\theta}_2^Z = Var\,\hat{\theta}_2^Z = \frac{2}{n - r(X)} \cdot \frac{[Tr(G)\mathbf{1}'\Sigma\mathbf{1}]^2 - \mathbf{2}Tr(G)\mathbf{1}'G\mathbf{1}\mathbf{1}'\Sigma^2\mathbf{1} + (\mathbf{1}'G\mathbf{1})^2 Tr(\Sigma^2)}{[(\mathbf{1}'w)^2 Tr(G) - \mathbf{1}'G\mathbf{1} w'w]^2}, \tag{12}$$

$$MSE\,\hat{\theta}_1^{YW} = Var\,\hat{\theta}_1^{YW} = \frac{2}{n - r(X)} \cdot \frac{1}{(p-1)^2 (w'G^{-1}w)^4} [(w'G^{-1}w)^2 Tr(G^{-1}\Sigma G^{-1}\Sigma) + \\ + (w'G^{-1}\Sigma G^{-1}w)^2 - 2(w'G^{-1}w)(w'G^{-1}\Sigma G^{-1}\Sigma G^{-1}w)], \tag{13}$$

$$MSE\,\hat{\theta}_2^{YW} = Var\,\hat{\theta}_2^{YW} = \frac{2}{n - r(X)} \cdot \frac{1}{(p-1)^2 (w'G^{-1}w)^4} [(w'G^{-1}w)^2 Tr(G^{-1}\Sigma G^{-1}\Sigma) + \\ + p^2 (w'G^{-1}\Sigma G^{-1}w)^2 - 2p(w'G^{-1}w)(w'G^{-1}\Sigma G^{-1}\Sigma G^{-1}w)]. \tag{14}$$

Keďže chceme porovnať tieto odhady na základe ich stredných kvadratických chýb, definujme si nasledujucúce funkcie

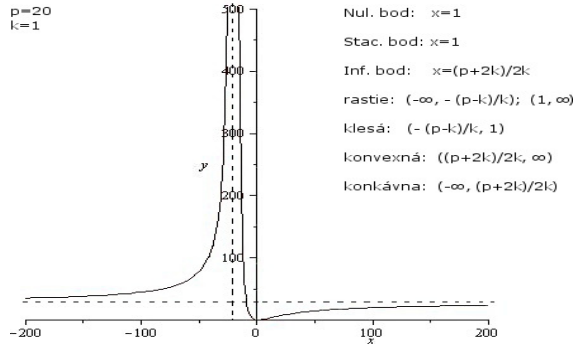
$$r_1 = MSE\,\hat{\theta}_1^Z - MSE\,\hat{\theta}_1^{YW} \quad \text{a} \quad r_2 = MSE\,\hat{\theta}_2^Z - MSE\,\hat{\theta}_2^{YW}. \tag{15}$$

4.1. Model zovšeobecnenej korelačnej štruktúry pre $\mathbf{w} = \mathbf{1}$

V prvom prípade budeme pracovať s jednotkovým vektorom *w* (*w*=**1**). Uvažujme teraz nasledujúcu štruktúru matice G. Nech má nenulové len diagonálne prvky, konkrétne nech obsahuje k premenných x a $(p - k)$ jednotiek. Teda je tvaru

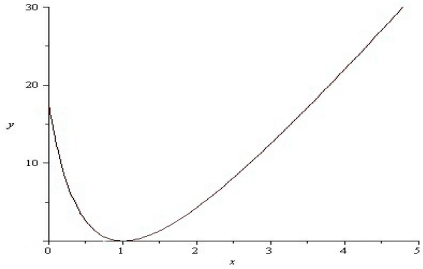
$$G = \mathrm{diag}(x, \ldots, x, 1, \ldots, 1).$$

Predpokladajme, že oba parametre θ_1 a θ_2 sú nezáporné. Skúmaním funkcií $r_1(x)$ a $r_2(x)$ dostaneme ich priebehy a vlastnosti zachytené v nasledujúcich obrázkoch.



Obrázok 1: Priebeh funkcie $r_1(x)$ pre w=1 a G=diag(x, … x, 1, …,1), $\theta_1 \geq 0$ a $\theta_2 \geq 0$.

Kedže G je diagonálna matica s diagonálnymi prvkami x a 1, tak aby Σ bola pozitívne semidefinitná, musí byť $x \geq 0$. Teda funkcie $r_1(x)$ a $r_2(x)$ stačí vyšetrovať na intervale $\langle 0, \infty)$. Na tomto intervale je $r_1(x) \geq 0$ pre každé $x \in R$. To znamená, že MSE $\hat{\theta}_1^Z \geq$ MSE $\hat{\theta}_1^{YW}$, takže v tomto prípade je odhad parametra θ_1 odvodený Ye a Wangom lepší ako odhad odvodený Žežulom. Rovnako funkcia $r_2(x)$ na intervale $\langle 0, \infty)$ vyzerá nasledovne



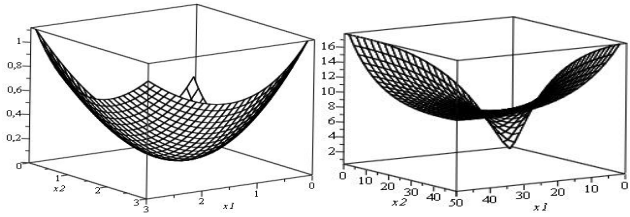
Obrázok 2: Priebeh funkcie $r_2(x)$ pre w=1, G=diag(x, …x, 1, …,1), $\theta_1 \geq 0$ a $\theta_2 \geq 0$.

Takže Ye a Wangov odhad parametra θ_2 v danom prípade je tiež lepší ako odhad odvodený Žežulom.

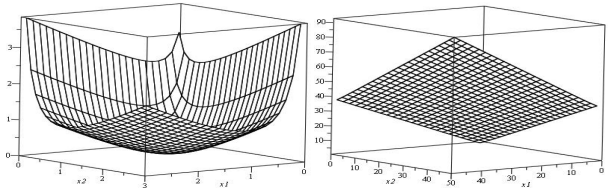
Uvažujme teraz prípad, kedy parametre θ_1 a θ_2 sú nezáporné a matica G obsahuje dve premenné. Teda je tvaru

$$G = \mathrm{diag}(x_1, \ldots, x_1, x_2, \ldots, x_2, 1, \ldots, 1).$$

Potom plochy určené funkciami $r_1(x_1, x_2)$ a $r_2(x_1, x_2)$ vyzerajú nasledovne



Obrázok 3: Plocha určená funkciou $r_1(x_1, x_2)$.



Obrázok 4: Plocha určená funkciou $r_2(x_1, x_2)$.

Aj v tomto prípade sú odhady Ye a Wanga pre parametre θ_1 a θ_2 lepšie, kedže plochy kriviek určených funkciami $r_1(x_1, x_2)$ a $r_2(x_1, x_2)$ ležia v nezápornej časti.

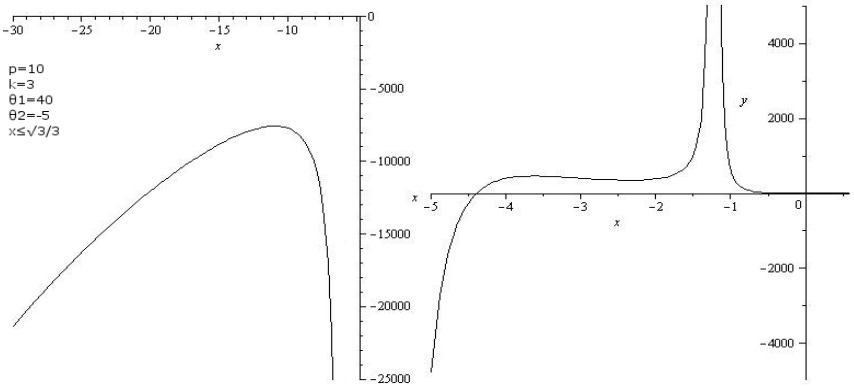
Ak by sme uvažovali parametre $\theta_1 \geq 0$ a $\theta_2 \leq 0$, tak v prípade, že $w = \mathbf{1}$ a $G = \mathrm{diag}(x, \ldots, x, 1, \ldots, 1)$ sa priebeh funkcie $r_1(x)$ výrazne nezmení. Takže odhad $\hat{\theta}_1^{YW}$ je lepší ako $\hat{\theta}_1^{Z}$. Lenže hodnoty funkcie $r_2(x)$ sú na niektorých podintervaloch intervalu $\langle 0, \infty)$ kladné a na niektorých záporné. Takže v tomto prípade je pre niektoré x lepší odhad $\hat{\theta}_2^{Z}$ ako $\hat{\theta}_2^{YW}$ a pre niektoré x to platí opačne. Určenie intervalov na ktorých je jeden odhad lepší ako druhý nie je jednoznačné. Rovnako to platí aj pre funkcie $r_1(x_1, x_2)$ a $r_2(x_1, x_2)$ v prípade, že $G = \mathrm{diag}(x_1, \ldots, x_1, x_2, \ldots, x_2, 1, \ldots, 1)$.

4.2. Model zovšeobecnenej korelačnej štruktúry pre $G = I$

Uvažujme teraz prípad, kedy G je jednotková matica a vektor *w* obsahuje k premenných x. Budeme teda pracovať s vektorom $w = (x, \ldots, x, 1, \ldots, 1)'$ a maticou $G = I$. Musíme však odvodiť podmienku semidefinitnosti matice Σ. V prípade, že $\theta_1 \geq 0$ a $\theta_2 \geq 0$, tak Σ je pozitívne semidefinitná pre každý vektor *w*. Takže x môže nadobúdať ľubovoľné hodnoty. V prípade, že $\theta_1 \geq 0$ a $\theta_2 \leq 0$, vychádzame z už uvedenej podmienky semidefinitnosti matice Σ, konkrétne musí platiť podmienka $|\theta_2| w' G^{+} w \leq \theta_1$. Čitateľ si vie jednoduchým dosadením odvodiť, že v tomto prípade je

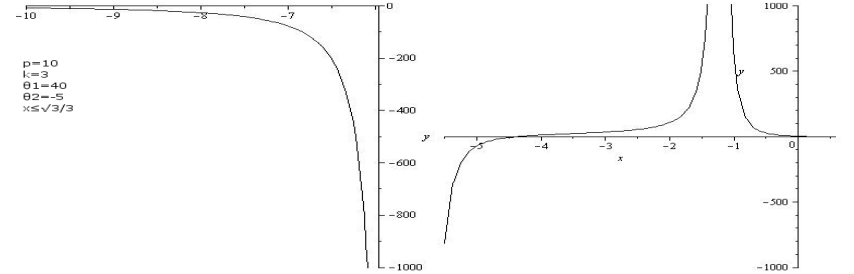
$$x \leq \sqrt{\frac{\theta_1}{k|\theta_2|} - \frac{p}{k} + 1}\,. \qquad (16)$$

Aby bola táto odmocnina definovaná, musí byť $\theta_1 \geq |\theta_2|(p-k)$. Priebeh funkcií r1(x) a r2(x) pre tento prípad je naznačený na nasledujúcich obrázkoch.



Obrázok 5: Priebeh funkcie r1(x) pre G=I, w=(x,…,x,1,…,1)′, $\boldsymbol{\theta_1 \geq 0}$ ***a*** $\boldsymbol{\theta_2 \leq 0}$.

Na nasledujúcom obrázku znázorníme priebeh funkcie $r_2(x)$ pre ten istý prípad a s rovnakými parametrami. Grafy týchto funkcií vyzerajú dosť podobne, lenže všimnime si rozdielnu mierku y-ovej osi. Z týchto grafov tiež vyplýva, že pre klesajúce x je v tomto prípade oveľa lepší odhad $\hat{\theta}_1^Z$ ako $\hat{\theta}_1^{YW}$, pretože $\left(\text{MSE}\,\hat{\theta}_1^{\text{Z}} - \text{MSE}\,\hat{\theta}_1^{\text{YW}}\right) \to -\infty$. Odhady parametra θ_2 sú pre klesajúce x v tomto prípade takmer rovnaké, kedže ako vidieť na nasledujúcom obrázku, rozdiel stredných kvadratických chýb pre klesajúce x konverguje k nule.



Obrázok 6: Priebeh funkcie r2(x) pre G=I, w=(x, …, x, 1, …, 1)′, $\boldsymbol{\theta_1 \geq 0}$ ***a*** $\boldsymbol{\theta_2 \leq 0}$.

Obe tieto funkcie nadobúdajú na niektorých podintervaloch intervalu $(-\infty, \sqrt{\frac{\theta_1}{k|\theta_2|} - \frac{p}{k} + 1}\rangle$ kladné a na niektorých záporné hodnoty. Takže na niektorých podintervaloch sú lepšie odhady odvodené Žežulom a na niektorých odhady odvodené Ye a Wangom. Presné určenie týchto intervalov je však aj v tomto prípade nejednoznačné. Rovnaké výsledky dostaneme aj v prípade, keď vektor w je tvaru $w = (x_1, \ldots, x_1, x_2, \ldots, x_2, 1, \ldots, 1)'$. Podmienka semidefinitnosti matice Σ bude teraz

$$x_1 \leq \sqrt{\frac{\theta_1}{k_1|\theta_2|} - \frac{k_2 x_2{}^2}{k_1} - \left(\frac{p}{k_1} - \frac{k_2}{k_1} - 1\right)}. \qquad (17)$$

Aby bola táto odmocnina definovaná, musí byť $\theta_1 \geq |\theta_2|[k_2 x_2{}^2 + (p - k_1 - k_2)]$.

Ak uvažujeme $\theta_1 \geq 0$ a $\theta_2 \geq 0$, tak za týchto podmienok nadobúdajú funkcie r1(x) a r2(x) nezáporné hodnoty pre všetky $x \in \mathbf{R}$. Takže $\hat{\theta}_1^{YW}$ je lepší ako $\hat{\theta}_1^Z$ a $\hat{\theta}_2^{YW}$ je lepší ako $\hat{\theta}_2^Z$. Rovnako aj pre $w = (x_1, \ldots, x_1, x_2, \ldots, x_2, 1, \ldots, 1)'$.

5. Záver

V tomto článku sme porovnali odhady variančných parametrov v modeli rastových kriviek s rovnomernou korelačnou štruktúrou, ktoré boli odvodené Žežulom, s odhadmi odvodenými Ye a Wangom. Ukázali sme, v ktorých prípadoch sú lepšie Ye-Wangove a v ktorých Žežulove odhady. Poukázali sme aj na prípady nejednoznačnosti rozhodnutia, ktorý odhad je na ktorom intervale lepší a v niektorých prípadoch sme graficky znázornili priebeh vyššie definovaných funkcií r_1 a r_2.

6. Literatúra

[1] Žežula, I. 2006. Special variance structure in the growth curve model. In: Journal of Multivariate Analysis, č. 97, 2006, 606 – 618.

[2] Ye, Ren-Dao – Wang, Song-Gui. 2009. Estimating parameters in extended growth curve model with special covariance structures. In: Journal of Statistical Planning and Inference, č. 139, 2009, 2746 – 2756.

Adresa autora:

Rastislav Rusnačko, Mgr.
Jesenná 5
041 54 Košice
rasto.rusnacko@gmail.com

Algoritmus DBSCAN a jeho realizácia v jazyku R
DBSCAN algorithm in R language

Miroslav Sabo

Abstract: Density-based clustering methods are one of the newest approaches in cluster analysis. One of them, DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is very popular especially for its simplicity and very good time-complexity. This submission explains how it works and shows practical realization in R language.

Key words: DBSCAN, clusters.

Kľúčové slová: DBSCAN, zhluky.

JEL classification: C38.

1. Úvod

Hlavnou myšlienkou algoritmov založených na hustote je predstava, že v okolí niektorých objektov vo viacrozmernom priestore sa nachádza viacero ďalších objektov. A naopak, niektoré objekty sú pomerne osamotené a nemajú dostatočný počet „susedov". Z toho môžeme povedať, že v okolí niektorých objektov je vyššia hustota iných objektov . Je intuitívne zrejmé, že pomocou tejto hustoty by sme mohli definovať jednotlivé zhluky ako oblasti s vyššou hustotou objektov. Analogicky, outliery sú zase objekty, v ktorých okolí je veľmi nízka hustota ďalších objektov. Najznámejším a najpoužívanejším algoritmom je DBSCAN, prvýkrát spomenutý v roku 1996 v [2]. V súčasnosti už existujú aj niektoré jeho zovšeobecnenia, ale aj úplne nové prístupy v algoritmoch založených na hustote.

2. Popis algoritmu DBSCAN

Detailný popis algoritmu možno nájsť napríklad v [1] alebo v [2]. Pred popisom samotného algoritmu definujme najskôr niektoré kľúčové pojmy. Všetky z nich pritom myslíme v zmysle hustoty bodov.

Definícia 1: $\varepsilon-$okolie bodu $x \in D$ je množina $N_\varepsilon(x) \equiv \langle y \in D : d(x, y) \leq \varepsilon \rangle$

Ide teda o všetky objekty z datasetu (označeného ako D), ktorých vzdialenosť od konkrétneho objektu je menšia alebo rovná ako nejaká predpísaná hodnota ε. O tom ako sa táto hodnota stanoví budeme hovoriť neskôr. Funkcia $d(x, y)$ je pritom nejaká miera vzdialenosti. Všimnime si, že do $\varepsilon-$okolia bodu patrí aj bod sám.

Definícia 2: Hovoríme, že bod x je priamo dosiahnuteľný z bodu y vzhľadom na dané ε a $N_{\min}$ ak platí

a) $x \in N_\varepsilon(y)$

b) $\left|N_\varepsilon(y)\right| \geq N_{\min}$

Inými slovami, podmienka a) vraví, že bod x sa musí nachádzať dostatočne blízko bodu y. Podmienka b) sa týka len bodu y a je to akési kritérium bodu s vyššou hustotou. Daný objekt teda nemusí byť priamo dosiahnuteľný z hocijakého bodu, ale len z takého, ktorý má vo

svojom okolí dostatočne veľa bodov. Symbolom $|N_\varepsilon(y)|$ pritom označujeme počet prvkov danej množiny a symbol $N_{\min}$je vopred stanovená hodnota podobne ako ε. O jej odhadnutí budeme rozprávať neskôr. Všimnime si ďalej, že ak je bod x priamo dosiahnuteľný z bodu y, tak automaticky neplatí, že bod y je tiež priamo dosiahnuteľný z bodu x. Symetria teda všeobecne neplatí. Jedná sa špeciálne o situáciu, kedy by bod x ležal na okraji zhluku a bod y v jeho vnútri, teda v mieste kde je vyššia hustota ako na okraji. Symetria zrejme platí, len ak sú oba body vnútri zhluku, teda sú pri sebe dostatočne blízko a zároveň v okolí každého z nich je dostatočný počet ďalších bodov. Všimnime si navyše, že každý bod nemusí byť priamo dosiahnuteľný sám zo seba. Definíciu 2 zrejme môžeme intuitívne rozšíriť nasledovne

Definícia 3: Bod x je dosiahnuteľný z bodu y ak $\exists$ postupnosť bodov $x_1 = x\ x_2 \ldots x_i = y$ taká, že x_k je priamo dosiahnuteľný z bodu x_{k+1} pre $k = 1,2,\ldots,i-1$. Opäť vzhľadom na dané ε a $N_{\min}$.

Dosiahnuteľnosť sme definovali pomocou priamej dosiahnuteľnosti. Dosiahnuteľnosť nie je symetrická, je iba tranzitívna.

Definícia 4: Hovoríme, že 2 body x a y sú prepojené vzhľadom na dané ε a $N_{\min}$ ak $\exists$ taký bod z, že oba body sú z tohto bodu dosiahnuteľné.

Definícia 5: Zhlukom C vzhľadom na dané ε a $N_{\min}$je neprázdna podmnožina D taká, že platí

a) ak $x \in C$ a y je dosiahnuteľný z x, tak potom aj $y \in C$.

b) $\forall x, y \in C$ platí, že x a y sú prepojené.

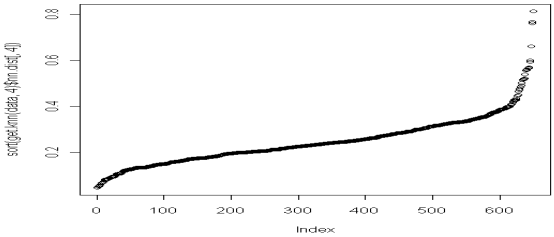
Prvá podmienka nám teda dáva priamo akýsi návod ako skonštruovať zhluk. Ak budeme mať nejaký konkrétny objekt, ktorý do neho patrí, tak automaticky k nemu pripojíme aj všetky ďalšie objekty, ktoré budú z neho dosiahnuteľné. Každý bod datasetu môžeme jednoznačne charakterizovať buď ako vnútorný, okrajový alebo vonkajší bod (outlier).

a)Vnútorné body - všetky také body, čo majú vo svojom $\in -$ okolí aspoň $N_{\min}$ bodov

b)Okrajové body - všetky také body, ktoré síce nemajú vo svojom okolí dostatočne veľa bodov, ale sú v blízkom okolí nejakého vnútorného bodu

c)Outliery - všetky ostatné body, teda body, okolo ktorých je veľmi málo iných bodov a pri ktorých nie je ani žiadny vnútorný bod

Pozrime sa teraz na stanovenie dvoch neznámych parametrov ε a $N_{\min}$. Treba pitom poznamenať, že ich určenie je sčasti subjektívne a musíme pri ňom často zohľadniť aj samotnú povahu dát. Definujme najskôr funkciu $D_k(x)$ nasledovne: $D_k(x)$ = vzdialenosť objektu x od svojho $k-$teho najbližšieho suseda (podľa nejakej metriky). Zistime teraz hodnoty tejto funkcie pre všetky objekty z datasetu a usporiadajme tieto hodnoty podľa veľkosti od najmenšej po najväčšiu. Ak tieto hodnoty nanesieme na dvojrozmerný graf, uvidíme podobný obrázok



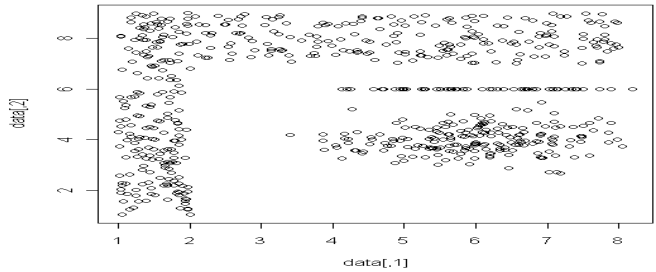
Obrázok 1: Funkcia štvrtého najbližšieho suseda

Z obrázka je na začiatku vidieť veľmi blízke objekty. Za nimi nasleduje dlhý úsek, ktorý veľmi pomaly rastie. Na konci nastáva niekoľko zlomových objektov, za ktorými už graf rastie pomerne výrazne. Práve tento zlom je dôležitý na stanovenie parametra ε. Za túto hodnotu teda vezmeme takú y – ovú súradnicu grafu, od ktorej začína graf prudko narastať. Skúmaním vlastností funkcie D_k sa navyše zistilo, že pre hodnoty k väčšie ako 4 sa vzhľad grafu príliš nemení. Odporúča sa teda skúmať len graf funkcie D_4. Z tohto dôvodu sa aj hodnota druhého parametra, teda $N_{\min}$ stanovila na 4.

3. Realizácia algoritmu v prostredí jazyka R

V jazyku R teraz ukážeme konkrétnu realizáciu algoritmu. Samotný programovací jazyk je pritom voľne stiahnuteľný a už má zabudovaný algoritmus DBSCAN, ktorý nájdeme v balíku „fpc“.

Pomocou generátora pseudonáhodných čísel v jazyku R vytvoríme dataset. Za účelom názornosti použijem iba 2 premenné, aby sa objekty dali vizualizovať v 2D priestore. Cieľom je zistiť, ako si poradí DBSCAN s našimi dátami, pričom budeme vopred z 2D grafu vidieť približné rozdelenie do zhlukov. Vygenerujme 3 zhluky. Prvý v tvare písmena „L“ a to z rovnomerného rozdelenia. Druhý v tvare priamky z normálneho rozdelenia. A tretí v tvare kruhu z dvoch normálnych rozdelení. Týmto chceme ukázať, že DBSCAN si poradí nielen s rôznymi tvarmi, ale často aj s rôznymi rozdeleniami. Predpokladom je len aby tieto rozdelenia nemali príliš odlišné smerodajné odchýlky. Dataset vygenerujeme pritom tak, aby sa dané 3 zhluky príliš neprekrývali. Spomínali sme totiž, že v takom prípade by DBSCAN nemusel priniesť želané výsledky.



Obrázok 2

Dodávame ešte, že prvý zhluk má 400 objektov, druhý 50 a tretí 200 (Obrázok 2). Zhluky sa teda významne líšia aj počtom objektov. Cieľom je teraz stanoviť hodnotu parametra ε v zmysle predošlích úvah. Za týmto účelom musíme zobraziť funkciu švrtého najbližšieho suseda a to príkazom plot(sort(get.knn(data,4)$nn.dist[,4])). Z jej grafu (obrázok 1) následne identifujeme bod, v ktorom graf začína prudko rásť. Stanovíme $\varepsilon = 0{,}4$. Minimálny počet bodov v zhluku je v našom príklade zbytočný, keďže zhluky majú veľa objektov. Počítajme ale s $N_{\min} = 5$, keďže program dbscan() má túto hodnotu už prednastavenú. Zadaním príkazu dbscan(data,0.4,showplot=0) dostaneme výstup, ktorý je v skrátenej verzii v tabuľke 1.

Vidíme, že DBSCAN správne určil počet zhlukov teda 3. Rovnako každý objekt, ktorý zaradil do niektorého zhluku do tohto zhluku aj skutočne patrí. Program určil aj 14 outlierov. Výstupom programu je aj presné zaradenie každého objektu medzi zhluky a počty vnútorných a vonkajších bodov v daných zhlukoch ako aj počty outlierov. Dodávame, že slovom seed sú označené vnútorné body. Pri zaradení do zhlukov sú zase outliery kódované číslom 0, a číslami 1,2,3 sa myslí príslušnosť objektu do daného zhluku.

Tabuľka 1: Výstup programu dbscan() v skrátenej verzii

```
dbscan Pts=650 MinPts=5 eps=0.4
        0   1  2   3
seed    0 387 45 181
border 14  11  3   9
total  14 398 48 190
[1]   1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[408] 2 2 2 2 0 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[630] 3 3 0 3 3 3 3 3 3 3 3 3 3 3 3 3 0 0 3 3 3 3
```

4. Záver

Výhodnosť algoritmu DBSCAN sa ukazuje hlavne v prípade väčších datasetov a pri zhlukoch rôznych tvarov. V prostredí jazyka R sa navyše realizuje veľmi jednoducho a názorne. Jeho nevýhodou však môžu byť zhluky s rôznymi hustotami objektov.

5. Literatúra

[1] Gan, Guojun, Chaoqun Ma, and Jianhong Wu, Data clustering: Theory, Algorithms and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007 219-227.

[2] Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, The Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996.

Adresa autora:

Miroslav Sabo, Mgr.
Radlinského 11
81368 Bratislava
miroslav.sabo@stuba.sk

Poďakovanie: *Tento článok vznikol za podpory grantového projektu VEGA č.1/0103/10.*

Metodologické aspekty hodnotenia a testovania modelov
Methodological aspects of evaluation and testing models

Tatjana Šimanovská

Abstract: The subject of our discussion are methodological aspects of evaluation and testing of models, which are the preferred research tool in economics, management, planning and design of socio-economic systems. Global knowledge society requires new and creative solutions to complex problems. Virtual reality, visualization and virtual labs provide us with space for creation and testing our ideas, our ingenuity and intellect through computational technology.

Key words: practical sciences, computational technology, model, evaluation, testing, alternatives, variables, creativity, decision-making, statistics.

Kľúčové slová: praktické vedy, komputačná technológia, model, hodnotenie, testovanie, alternatívy, premenné, kreatívnosť, rozhodovanie, štatistika.

JEL classification: C18, C19, C51, C52, C88

1. Úvod

Predmetom našich úvah budú niektoré metodologické aspekty hodnotenia a testovania modelov, ktoré patria k preferovaným výskumným nástrojom v globálnej vedomostnej spoločnosti, ktorá vyžaduje nové kreatívne riešenia komplexných problémov. Rozvoj technológií v oblasti IKT pokračuje závratným tempom a ponúka množstvo nekonvenčných postupov a produktov, v ktorých je už aj náročné sa orientovať. Virtuálna realita, vizualizácia a virtuálne laboratóriá nám poskytujú priestor pre kreovanie a testovanie našich ideí, našej invencie a intelektu prostredníctvom komputačných technológií..

2. Hodnotenie a testovanie modelov

Hodnotenie je dôležitou fázou každého technologického procesu. M. Bunge zdôrazňuje význam hodnotenia vo všetkých fázach technologickej činnosti, od jej prípravy a predpokladov, až po jej dôsledky. Bohužiaľ sa často stáva, že hodnotenie sa vždy nevykonáva a, čo je veľmi dôležité, nevykonáva sa vždy vedecky. Vyzdvihuje úlohu experimentu, ktorý môže s určitosťou dokázať či nový artefakt prináša úžitok alebo škodu. Experiment nemôžu nahradiť ani dobré úmysly, ani intuície, a preto spoločnosť musí rozvíjať vedeckú technológiu. (Pozri [4], s. 231.) V tejto súvislosti by sme radi zdôraznili myšlienku F. Novosada, že opatrnosti nie je nikdy dosť a dobré úmysly sú to, čo sa v dejinách najmenej ráta. (Pozri [7], s. 79.)

Aj podľa A. Podgóreckého hrá hodnotenie dôležitú rolu v praktických disciplínach a hodnotenia nepripravujú praktické vedy o objektívny status. Možno rozlíšiť dva typy hodnotení: vlastné hodnotenia a utilitaristické hodnotenia. Utilitaristické hodnotenia možno klasifikovať podľa troch kritérií: a) vzhľadom na ich termíny alebo spôsob vyjadrovania; b) vzhľadom na ich záujem (prostriedky, výsledky); c) vzhľadom na ich miesto v systéme hodnôt. V diskusii o role hodnotenia v praktických vedách je podstatné indikovať nielen v ktorej fáze procedúry sa hodnotenie vyskytuje, ale aj v akej forme. (Pozri [8], s. 91 - 92.) Vlastné hodnotenia sa vyskytujú v podstate len vo fáze výskumnej procedúry v praktických vedách, t.j. vo fáze kritického triedenia hodnotení prijateľných v sledovaní danej praktickej akcie. Avšak tieto hodnotenia nie sú časťou jazyka praktických vied. Vo všetkých ostatných

fázach praktickej procedúry sú utilitaristické súdy. Tieto súdy sú intersubjektívne a nezavádzajú žiadne elementy subjektívnej svojvôle do praktických disciplín. Inými slovami hodnotenia v praktických vedách vznikajú buď ako objektívne dáta zvonku, alebo ako utilitaristické súdy bez subjektívnych elementov. (Pozri tamže, s. 93.) Poznamenávame, že pre hodnotenie dopadov a účinkov technológie sa vžil termín „technology assessment," pre ktorý L. Tondl presadzuje názov „sociálne hodnotenie techniky" a napísal aj rovnomennú prácu [9]. Táto oblasť skúmania sa v súčasnosti mimoriadne rozvíja, najmä v dôsledku globalizácie problémov a rizík zavádzania výsledkov technologickej činnosti.

Významnou súčasťou procesu plánovania je aj hodnotenie modelov. Spoľahlivosť hodnotení prostriedkov (napr. činy, postupy činnosti, praktiky, projekty, atď.) založená na modelovaní, závisí od toho, ako dobre modely reprezentujú realitu. Preto je dôležité hodnotiť samotné modely pred ich použitím na hodnotenie prostriedkov. Na hodnotenie modelov môžu byť efektívne štatistické techniky, ktoré sú zárukou spoľahlivosti modelov, a preto by ich plánovači mali dôkladne poznať.

Pri hodnotení prostriedkov možno obmedzene použiť nekompletné modely, v ktorých chýbajú relevantné premenné buď preto, že ich bolo ťažko kvantifikovať, alebo z iných dôvodov. V takomto prípade je významné, aby si rozhodujúci sa subjekt uvedomil, ktoré premenné sa v nekompletných modeloch brali do úvahy, a ktoré nie. Môže doplniť výstup modelu posúdením vylúčených premenných a dosiahnuť tak lepšie rozhodnutie. Preto je veľmi dôležité pre rozhodujúce sa subjekty vedieť nielen to, čo je a čo nie je zahrnuté v modeli, ktorého výsledok použijú, ale aj to, ako dobre tieto modely reprezentujú realitu. Pred úvahami o spôsobe hodnotenia modelov je vhodné uvažovať o možných spôsoboch použitia modelov na hodnotenie prostriedkov.

V tejto súvislosti si treba uvedomiť, že modely majú aj určité nedostatky. Nedostatky modelu sú rovnaké ako nedostatky rozhodovania, pretože model používaný na hodnotenie prostriedkov je model rozhodovania. Podľa R.L. Ackoffa môžu nedostatky vzniknúť z týchto dôvodov: 1) vynechanie relevantných premenných a/alebo zahrnutie irelevantných premenných; 2) zlyhanie riadenia ovplyvniteľnej premennej; 3) vynechanie relevantných obmedzení a/alebo zahrnutie irelevantných a 4) nesprávna formulácia vzťahu medzi premennými a výsledkami, ktoré môžu nastať ([1], s. 205).

G. Booch tvrdí, že modelovanie realizuje princípy dekompozície, abstrakcie a hierarchie. Modely nám dávajú možnosť skúmať nedostatky systému v podmienkach, ktoré sme si sami určili. Hodnotíme správanie každého modelu v bežných aj neobvyklých situáciách, a potom uskutočňujeme potrebné úpravy. Je efektívne používať viac modelov, aby sme pochopili všetky detaily zložitého systému. (Pozri [3], s. 28)

Konštrukcia modelov je v podstate generalizáciou procesu formulovania alternatívnych prostriedkov. Tento proces vyžaduje veľkú kreatívnosť. Je dôležité si uvedomiť, že predstava rozhodujúceho sa o situácii voľby sa môže úplne odlišovať od reálnej situácie. V prípade, že objavíme takéto rozdiely, otvára sa nám možnosť formulovať nové prostriedky. Rozdiely sa obyčajne skrývajú za predpokladmi, ktoré sa rozhodujúcemu sa zdajú také pravdivé, že o nich vôbec nepochybuje. Taký postoj však môže byť zdrojom budúcich, zdanlivo neriešiteľných problémov. Situácia rozhodovania je proces kladenia otázok a odpovedí na ne, ktorý vychádza často s protirečivých predpokladov. Napríklad otázka relevantnosti premenných v určitej situácii závisí od sústavy súradníc, alebo od hľadiska prístupu k situácii. Naša sústava súradníc v špecifickej situácii tvorí súčasť svetonázoru, ktorý je produktom našej kultúry, výchovy, povolania a záľub. Pretože je svetonázor zriedka jasne vyjadrený a málokedy sa hodnotí a reviduje, si často neuvedomujeme mieru, s akou vylučujeme premenné z úvah o špecifických situáciách rozhodovania. Našu sústavu súradníc si uvedomujeme prostredníctvom iných ľudí pri diskusiách o problémoch. Preto sa v súčasnosti kladie dôraz na

spoluprácu a pomoc iných ľudí s odlišnými hľadiskami, lebo čím je viac rozmanitých pohľadov na problém, tým sa vytvára viac alternatívnych ciest jeho riešenia.

Model tiež reprezentuje environment a špecifikuje, akým spôsobom na výsledok vplývajú zmeny akejkoľvek premennej, riadenej či neriadenej. V súvislosti s konštrukciou modelov je nevyhnutné uvažovať o ich testovateľnosti. Je veľmi dôležité, aby sa testovateľnosť modelu zohľadnila počas jeho konštruovania. Model, ktorý nemožno testovať je len domnienkou. Je aktom viery a nie vedy.

V súčasnosti sa mnoho koncepcií venuje problematike testovania modelov. Modely možno napríklad podľa R.L. Ackoffa testovať buď retrospektívne, alebo perspektívne. Retrospektívne testovanie šetrí čas a jeho obsahom je hľadanie alebo rekonštrukcia hodnôt riadených, neriadených a výsledkových premenných pre všetky periódy, ktoré sú z hľadiska modelovania dôležité. Potom hodnoty riadených a neriadených premenných umiestnime do modelu a snažíme sa odhadnúť výsledkovú premennú. Porovnanie sa obyčajne realizuje vo forme testovania hypotézy, v ktorej priemerný rozdiel medzi modelovo odhadovaným a aktuálnym fungovaním by mal byť nula. Taktiež sa hodnotí spoľahlivosť odhadov, t.j. ich rozptyl okolo aktuálneho fungovania. Avšak významným obmedzením retrospektívneho testovania modelov je nutnosť, aby sa budúca situácia veľmi neodlišovala od minulosti. Čím viac sa budúcnosť odlišuje od minulosti, tým menej vhodné je retrospektívne testovanie. Perspektívne testovanie má rovnakú logiku ako retrospektívne, s tým rozdielom, že hodnoty relevantných premenných sa pozorujú v budúcich obdobiach. V obidvoch typoch testovania je mimoriadne dôležité, aby sa skutočné výsledky hodnotili nezávisle od modelu, inak je test bezvýznamný. (Pozri [1], s. 206.)

V niektorých prípadoch sa model nedá testovať ani retrospektívne, ani perspektívne. V týchto prípadoch sa odporúča hodnotiť model prostredníctvom analýzy senzitivity, ktorá je metódou testovania citlivosti modelu na veľkosť vplyvu premenných na výsledky modelu. Táto analýza obsahuje určenie veľkosti chyby v odhadoch hodnôt premenných použitých v procese hodnotenia, ktorou najlepší prostriedok špecifikovaný modelom funguje menej uspokojivo, než jeho alternatíva. Použitie tejto metódy sa však obmedzuje len na modely s malým počtom premenných.

R.L. Ackoff na základe vzťahov medzi rozhodovacími premennými, ktoré môžu byť buď asociácie - čisto deskriptívne vzťahy, alebo vzťahy príčina a účinok a výrobca a produkt - explanačné vzťahy, rozlišuje *deskriptívne a explanačné modely*. Deskriptívne modely sa zakladajú na asociáciách a explanačné modely vychádzajú zo vzťahov príčina a účinok a výrobca a produkt. Deskriptívne modely opisujú len to, ako sa hodnoty rôznych premenných navzájom menia a neopisujú, ako zmeny jednej premennej zapríčiňujú, alebo produkujú zmenu inej premennej. Explanačný model reprezentuje spôsob, akým jedna alebo viac premenných ovplyvňuje, alebo produkuje zmeny jednej, alebo viac výsledných premenných. Preto ho možno použiť na hodnotenie prostriedkov. (Pozri [1], s. 207 - 208.)

Deskriptívne modely určitého typu možno použiť pri predikcii. Sú to modely, v ktorých sa hodnoty niektorých premenných v jednom bode času spájajú s hodnotami iných premenných v neskoršom bode času (napr. ekonometrické modely). Keďže implementácia prostriedkov obsahuje zmenu hodnôt riadených premenných a určovanie účinkov na jednu alebo viac výstupných premenných, nemožno deskriptívne modely použiť na hodnotenie prostriedkov, lebo prostriedok je riadená premenná, a preto výstup vyžaduje kauzálny vzťah. Deskriptívne modely sa často nesprávne a nevhodne používajú na hodnotenie prostriedkov. Ekonometrické modely sa zas používajú na hodnotenie alternatívnych politík a stratégií. Preto sa nemožno čudovať, že mnohé predikcie účinkov zásahov do ekonomiky nedosiahnu cieľ.

Mimoriadne inšpiratívne idey prináša L. Andrášik vo svojej práci [2]. Predstavuje možnosti, ktoré ponúkajú Systémová dynamika, Komplexita a Synergetika za pomoci IKT, Internetu, aplikovanej informatiky, komputačnej inteligencie a ďalších prostriedkov pri riešení problémov a menovitých úloh v oblasti ekonómie, manažmentu a plánovania. Moderná veda vyžaduje nekonvenčné metódy a postupy, založené predovšetkým na

vizualizácii vysoko abstraktných javov a procesov pomocou experimentovania vo virtuálnych laboratóriách. Autor používa prístupy a metódy, ktoré nespočívajú na verbálnom vyjadrovaní, ale na aktívnom zúčastňovaní sa na procesoch prebiehajúcich vo virtuálnej realite. Takto sa darí integrovať do harmonickej jednoty inštruktívne a konštruktívne prístupy k získavaniu nových poznatkov a zároveň zručností pre takéto špecifické činnosti, najmä pri tvorbe mentálnych modelov a pri ich konštrukčnej realizácii vo vhodných softvéroch, konkrétne v tabuľkovom procesore Excel, komerčnom softvéri STELLA a iDmc. (Pozri [2] , s. 299)

4. Záver

V eseji sme sa zamerali na metodologické aspekty hodnotenia a testovania modelov. Možno konštatovať, že užitočnosť modelov pri hodnotení prostriedkov závisí od toho, či modely opisujú alebo vysvetľujú relevantný jav. Deskriptívne modely možno použiť iba na predpovedanie toho, čo sa stane v prípade, že sa neplánuje žiadny zásah. Keďže prostriedky sú plánované zásahy, nemožno deskriptívne modely použiť na ich hodnotenie. Projekt efektívneho hodnotenia prostriedkov vyžaduje technickú zručnosť v konštruovaní modelov, v matematickom manipulovaní s nimi, v projektovaní prírodných a simulovaných experimentov a vo vykonávaní štatistických analýz dát, ktoré sa získali z experimentov obidvoch typov. Dobre riadené hodnotenie prostriedkov môže často navrhnúť spôsob formulovania nových prostriedkov, ktoré sú lepšie, než ktorékoľvek predtým testované. Navyše môže navrhnúť, ako formulovať prostriedky, ktoré sa používaním môžu zdokonaľovať, t.j. uľahčovať učenie sa a adaptáciu socioekonomických systémov v globálnej vedomostnej spoločnosti.

5. Literatúra

[3]ACKOFF, R. L.1981. Creating The Corporate Future. New York 1981.
[4]ANDRÁŠIK, L. 2010. Aplikovaná systémová dynamika a synergetika. STU v Bratislave 2010. 311 s. ISBN 978-80-227-3360 -1.
[5] BOOCH, G. 1992. Object Oriented Design with Applications. V ruskom preklade: Objektno orientirovannoje projektirovanie s primerami primenenija. Kyjev - Moskva 1992.
[6] BUNGE, M. 1985. Treatise on Basic Philosophy. Vol. 7, Part II. Dordrecht - Boston 1985.
[7] HOLOMEK, J., ŠIMANOVSKÁ T. 2002. Úvod do metodológie praktických vied. Bratislava 2002.163 s. ISBN 80-8054-249-X.
[8] CHAJDIAK. J. 2009. Štatistika v exceli 2007. Bratislava Statis 2009, ISBN 978-80-85659-49-8.
[9] NOVOSAD, F.1994. Vysvetľovanie rukami. IRIS, Bratislava 1994.
[10] PODGÓRECKI, A. 1975. Practical Social Sciences. Routledge & Kegan, London and Boston 1975.
[11] TONDL, L.:1992. Sociální hodnocení techniky. Plzeň 1992.

Adresa autora:

Tatjana Šimanovská, PhDr., PhD.
ÚM STU, OEMP
Vazovova 5, 812 43, Bratislava
tatjana.simanovska@stuba.sk

Príspevok bol spracovaný v rámci riešenia úlohy VEGA č. 1/0536/10 Inovácie ako strategický základ konkurenčnej schopnosti SR (Smerovanie, meranie a podpora inovačných procesov).

Aktuálny vývoj príjmovej nerovnosti na Slovensku
Recent trends in income inequality in Slovakia

Juraj Sipko, Ľubica Sipková

Abstract: With the current processes of globalization and deregulation, we may observe a relatively strong polarisation of incomes taking place in all countries worldwide, including European countries. A characteristic trait of the process of polarization is the steep rise in the incomes of the wealthy and the gradual decrease of incomes in the middle class. The aim was to find statistical measures, along with the corresponding graphical presentations, which would appropriately describe the shift to lower incomes in the middle class in recent years in the Slovak Republic. On the basis of the presented results, we conclude that the middle class is experiencing a proportional shift to lower incomes in the SR.

Key words: Social polarization, Income inequality, Kernel estimation, Gross wages, Income, Employees, Middle class, Quantiles, Galton's skewness.

Kľúčové slová: sociálna polarizácia, príjmová nerovnosť, kernelový odhad, hrubá mzda, príjem, zamestnanci, stredná vrstva, kvantily, Galtonova šikmosť.

JEL classification: J31, C14, C46.

1. Úvod

V súčasnosti pokračuje proces sociálnej polarizácie v každej z európskych krajín. Tento proces je regionálne nerovnomerný. V hodnotiacich správach o ekonomickom vývoji sa problematike sociálnej situácie stredne zarábajúcich vrstiev v krajinách EÚ zatiaľ nevenuje primeraná pozornosť. Pozornosť sociológov, regionálnych geografov, štatistikov, ekonómov a politikov v hodnoteniach o sociálnom vývoji v krajinách EÚ bola sústredená hlavne na najchudobnejšie skupiny obyvateľstva. Analýzy boli orientované hlavne na rozsah chudoby, riziko ohrozenia chudobou, rozsah sociálneho vylúčenia a na životné podmienky jednotlivcov nachádzajúcich sa na okraji spoločnosti.

Stredne zarábajúca vrstva (stredná vrstva) by mala predstavovať prevažnú časť v spoločnosti. Základom pre klasifikáciu strednej vrstvy je teda úroveň príjmov. Od úrovne príjmov strednej vrstvy obyvateľstva závisí aj jej kúpyschopný dopyt. Stredná vrstva je tou časťou spoločnosti, ktorá výraznou mierou ovplyvňuje agregovaný dopyt. Čím vyššie sú príjmy strednej vrstvy, tým je vyšší agregovaný dopyt a tým vyšší je hospodársky rast a zamestnanosť. Vyššie príjmy spravidla nielen podporujú hospodársky rast a znižujú nezamestnanosť, ale pozitívne ovplyvňujú aj sociálnu situáciu v spoločnosti a rast životnej úrovne obyvateľstva. Okrem toho, pre strednú vrstvu je charakteristická aj vzdelanostná úroveň a jej miesto v rozhodovacích procesoch v spoločnosti.

Na konci druhej polovice minulého storočia s prebiehajúcim procesom globalizácie, ale aj deregulácie možno sledovať, že dochádza k relatívne silnej polarizácii príjmov takmer vo všetkých priemyselne vyspelých, ale aj rozvojových štátoch, tzn. aj v európskych štátoch. Charakteristickou črtou nastúpeného procesu polarizácie je prudký nárast príjmov v skupine bohatých a postupné znižovanie príjmov v strednej vrstve. K samotnému procesu polarizácie v mnohých prípadoch prispievajú aj politické rozhodnutia v niektorých štátoch. Napr. stanovenie rovnakej dane z príjmov pre všetkých obyvateľov. Výsledkom procesu polarizácie je nespravodlivé prerozdeľovanie zdrojov v spoločnosti a tým ďalšie znižovanie príjmov strednej vrstvy.

Pri regionálnom NUTS-2 porovnaní jadrových (kernelových) odhadov funkcií hustôt (*pdfs*) rozdelení hrubých mesačných nominálnych miezd v roku 2007 sa potvrdila chýbajúca stredne zarábajúca vrstva zamestnancov v SR. Podobné presunutie pravdepodobnosti výskytu miezd zo strednej časti rozdelenia do jeho koncov bolo v každom kraji SR. Najväčší prepad strednej časti kernelového odhadu *pdf* bol v Bratislavskom kraji. Za ním nasledoval Západoslovenský kraj a najmenej výrazný bol v Stredoslovenskom kraji.

Ani jeden z testovaných známych tvarov *pdf* nebolo možné uznať za vhodný model kvôli príliš veľkému presunutiu pravdepodobnosti zo strednej časti smerom ku koncom rozdelení. Vyplýva to aj zo zápornej asymetrie strednej časti celkovo kladne asymetrického rozdelenia miezd a neúmerne predĺženého tenkého pravého konca empirických rozdelení. Výsledky analýzy kernelových odhadov a aplikácie pravdepodobnostného modelovania rozdelení miezd v NUTS-2 regiónoch SR v roku 2007 [1] nás viedli k skúmaniu vývoja príjmov stredne zarábajúcej vrstvy zamestnancov v období rokov 2005 až 2009.

Príspevok je venovaný porovnaniu rozdelení ročných príjmov zo zamestnania (peňažných a nepeňažných) u zamestnancov SR podľa výsledkov oficiálneho zisťovania EU-SILC. V príspevku konštatujeme, že na Slovensku pokračuje proces proporcionálneho posunu stredne zarábajúcich vrstiev zamestnancov SR bližšie smerom k nízkym príjmom. Vzhľadom na obmedzený rozsah príspevku, prezentované sú len niektoré zaujímavé výsledky, ktoré potvrdzujú tento záver.

2. Vstupné údaje a použité metódy

V príspevku prezentované výsledky vychádzajú zo súborov údajov harmonizovaného ročného zisťovania Štatistika o príjmoch a životných podmienkach v EÚ (EU-SILC) za roky 2005, 2006 (UDB zo 6.9.2007), 2007 (UDB_20/08/2008), 2008 (UDB_10/09/2009) a 2009 (UDB_26/07/2010), poskytnutých Štatistickým úradom SR. Získané boli stratifikovaným náhodným výberom, formou zisťovania PAPI v domácnostiach v SR v súlade s platnými nariadeniami Európskej komisie ohľadom výberu, spôsobu zisťovania a definovania premenných. Informácie o osobách vo veku 16 rokov a viac sú sústredené do samostatného súboru osobných prierezových údajov, ktorý je označený ako P_súbor.
Primárne premenné v P_súbore sú rozdelené do piatich oblastí. Údaje o osobnom príjme predstavujú hodnoty pätnástich premenných (séria komponentov hrubých príjmov a séria komponentov čistých príjmov). Aplikované v príspevku sú premenné: SPY200G Hrubá mesačná mzda zo zamestnania – Slovenská premenná zo súboru EU-SILC 2007; PY010G – Peňažný príjem zo zamestnania alebo jemu blízky príjem a PY020G – Nepeňažný príjem zo zamestnania, ročná suma pre obe premenné zo súborov za roky 2005 až 2009. Spracovávané boli len nenulové hodnoty príjmov osôb zaradených do kategórie 3 (zamestnanec pracujúci za mzdu, plat, iný druh odmeny) premennej PL040 – Ekonomické postavenie v zamestnaní.

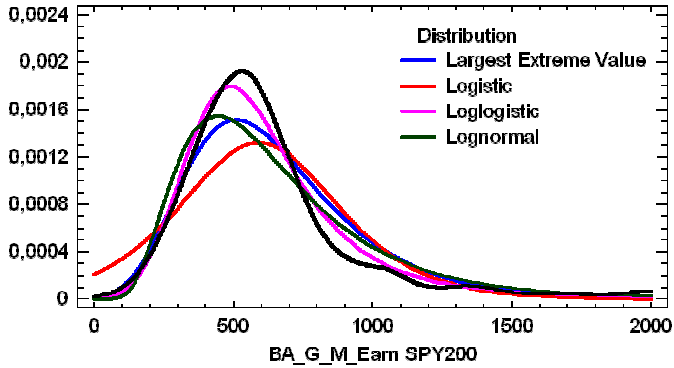
Treba podotknúť, že zistené údaje boli na ŠÚ SR kontrolované a upravované pre neodpovede, nekontaktovanie domácnosti, nevyčíslenie a zistené chyby. Pre príjmové premenné boli uplatnené imputácie, predovšetkým na základe priemerovania. Analyzované príjmové premenné, obsiahnuté v P_súbore, sú v príspevku označené v súlade s oficiálnym označením v databáze podľa nariadení EK a oficiálnych príloh k databáze, preto ich v článku nedefinujeme. Definície cieľových premenných možno nájsť na oficiálnej stránke EK v Nariadení Komisie (ES) č. 1983/2003 zo dňa 7. 11.2003. Kvôli rýchlejšej orientácii vo výstupoch a grafoch k oficiálnemu označeniu sme priradili krátke označenie s posledným dvojčíslom roka (napr. empl_05income).

Použité pre hodnotenie príjmovej nerovnosti a jej vývoja podľa vyššie uvedenej údajovej základne boli hlavne kvantilové deskriptívne a nadväzujúce induktívne štatistické metódy na báze teórie poriadkových štatistík. Analýzy boli robené v štatistických

programových balíkoch Statgraphics, SPSS a výpočty menej známych Galtonových kvantilových charakteristík šikmosti ako aj relatívnych mier nerovnosti pre volené časti rozdelení v Exceli 2007. Grafická analýza spočívala v porovnaní kernelových odhadov funkcií hustôt, Q-Q grafov empirických rozdelení poriadkových štatistík a box-plotov pre volené časti rozdelení. Zdrojom všetkých grafov a tabuliek prezentovaných v príspevku sú vlastné analýzy autorov.

3. Regionálna analýza mzdovej nerovnosti v SR v roku 2007

V grafickom porovnaní kernelového odhadu funkcie hustoty rozdelenia hrubých mesačných nominálnych miezd (najvýraznejšia krivka na obrázku 1) s najlepšími odhadmi známych tvarov *pdfs* možno vidieť rýchlejší pokles za vrcholom v oblasti hodnôt miezd od 650 € do 900 €. Naklonenie vrcholu v strednej časti rozdelenia upozorňuje na zápornú asymetriu strednej polovice hodnôt v rozdelení, ktoré má celkovo kladnú asymetriu. Treba si uvedomiť, že posun vrcholu smerom k vyšším hodnotám (záporná asymetria) len v strednej časti celkovo kladne zošikmeného rozdelenia neznamená nárast, ale práve prepad, posun pravdepodobnosti zo strednej časti rozdelenia do jeho koncov.



Obrázok 1: Kernelový odhad hrubých mesačných miezd zamestnancov v Bratislavskom regióne (s maximálne vierohodnými odhadmi teoretických rozdelení)

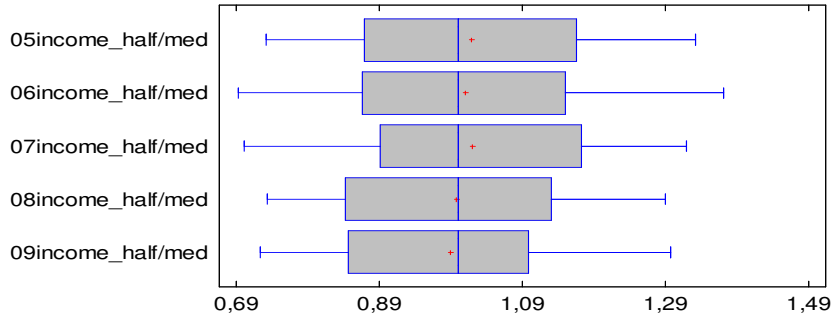
Pre pravdepodobnostné modelovanie takýchto netypických tvarov rozdelení miezd v regiónoch SR nebol nájdený vhodný známy tvar *pdf* z 38 testovaných teoretických spojitých rozdelení (najlepšie pre SR a jej regióny boli tvary loglogistického, lognormálneho, prípadne aj logistického rozdelenia). Overenie zhody rozdelení bolo uskutočnené viacerými známymi testami dobrej zhody, ale tak ako je časté v prípade veľkých výberov testy nepotvrdili dobrú zhodu. Preto sme porovnávali teoretické rozdelenia s empirickými pomocou Q-Q grafov a zhodu merali koeficientmi korelácie. Však ani jeden z testovaných tvarov nebolo možné uznať za vhodný kvôli príliš hrubým dolným koncom, zápornej asymetrii strednej časti a príliš predĺženému tenkému pravému koncu empirických rozdelení (výsledky regionálneho pravdepodobnostného modelovania premennej SPY200G v roku 2007 pozri napr. v [1]).

Ako dokumentuje dosiaľ uvedené, rozlíšiť možno dva prípady presunu pravdepodobnosti zo strednej časti rozdelenia miezd. Po prvé, polarizácia nastáva nárastom extrémne vysokých miezd a po druhé k polarizácii dochádza prepadom stredne zarábajúcej vrstvy smerom k nízkym hodnotám. Často ide o nerovnomernú kombináciu oboch prípadov. Presunom pravdepodobnosti do tenkého pravého konca, t. j. častejším výskytom extrémne vysokých miezd v SR sa v tomto príspevku nezaoberáme.

Sústredili sme pozornosť na nárast príjmovej nerovnosti, ktorý je spôsobený postupným presunom pravdepodobnosti v prostrednej časti rozdelenia smerom k dolnému (ľavému) hrubému koncu rozdelenia príjmov zamestnancov SR.

4. Posledný vývoj príjmovej nerovnosti v strednej vrstve zamestnancov SR

Strednú vrstvu zamestnancov definujeme podľa úrovne príjmov, t. j. patria do nej zamestnanci s príjmami medzi dolným a horným kvartilom príjmového rozdelenia. Rozsah rastu príjmovej nerovnosti len posunom v strednej časti rozdelenia príjmov zamestnancov smerom k nižším hodnotám v SR v poslednom období hodnotíme graficky, napr. porovnaním box-plotov len pre prostrednú časť, ale aj kvantitatívne, napr. kvantilovými charakteristikami asymetrie strednej časti rozdelenia miezd v SR v rokoch 2005 až 2009. Rozdelenie časti medzi dolným a horným kvartilom príjmovej premennej (peňažný aj nepeňažný) príjem zo zamestnania, ktorá vznikla súčtom PY020G a PY010G, je za roky 2005 až 2009 porovnaný na obrázku 2 pomocou box-plotov. Osoby zahrnuté do kategórie 3 (zamestnanec), ktoré mali nulové hodnoty sledovanej premennej, neboli do analýzy zahrnuté.



Obrázok 2: Box-ploty prostrednej časti relatívnych hodnôt príjmových rozdelení vzhľadom na ročné mediány v rokoch 2005 až 2009

Pretože sledovaný príjem je nominálnou premennou, rastúcou na Slovensku v rokoch 2005 až 2009, porovnania na základe absolútnych hodnôt príjmov vhodne necharakterizujú posuny v proporcionálnej štruktúre rozdelenia príjmov. Cieľom nie je hodnotiť vývoj úrovne príjmov, ale názorne porovnať proporcionálne zmeny v prostrednej časti rozdelenia, preto sme na obrázku 2 použili príjmy ustálené na mediánovú úroveň v sledovaných rokoch.

Tabuľka 1: Hodnoty Galtonovej šikmosti – kvantilové miery asymetrie prostrednej časti príjmových rozdelení, 2005 - 2009

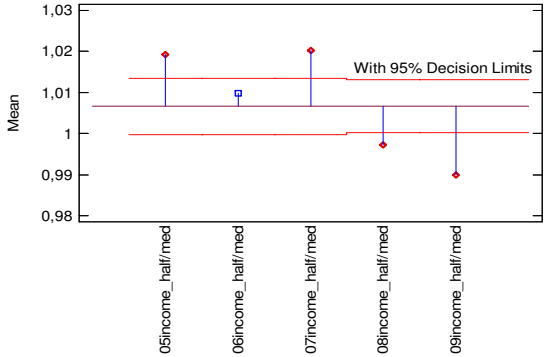
Príjmy medzi dolným a horným kvartilom	kvartilová šikmosť	sixtilová šikmosť	kvartilová šikmosť		sixtilová šikmosť	
			ľavá	pravá	ľavá	pravá
empl_05income_half	0,1169	0,2441	0,442	0,558	0,411	0,589
empl_06income_half	**0,0588**	0,0000	0,471	0,529	0,500	0,500
empl_07income_half	**0,2292**	0,2856	0,385	0,615	0,395	0,605
empl_08income_half	**-0,0913**	**-0,0521**	0,546	0,454	0,522	0,478
empl_09income_half	**-0,2194**	**-0,2294**	0,610	0,390	0,576	0,424

Zaujímavé je postupné skracovanie hornej časti škatuľky sprevádzané nárastom jej ľavej časti, posun celej škatuľky k nižším hodnotám, ako aj zmena polohy priemeru pod medián v roku 2008 a jeho ešte väčší posun k nižším hodnotám v roku 2009. Posun

priemeru a nárast ľavej časti škatuľky v porovnaní s pravou časťou znamenajú rast zápornej asymetrie prostrednej časti príjmových rozdelení, čo je v grafe znázornené aj proporcionálne v súlade s narastajúcou vzdialenosťou kvartilov relatívnych hodnôt príjmov.

Veľkosť tohto posunu v prostrednej časti príjmového rozdelenia možno merať pomocou relatívnej charakteristiky šikmosti na báze kvantilov, tzv. Galtonovej šikmosti. Pomocou sixtilovej šikmosti meriame asymetriu väčšej časti prostrednej polovice príjmového rozdelenia porovnaním štandardizovaných vzdialeností horného sixtilu od mediánu a dolného sixtilu od mediánu. Výsledné hodnoty Galtonových mier obsahuje tabuľka 1.

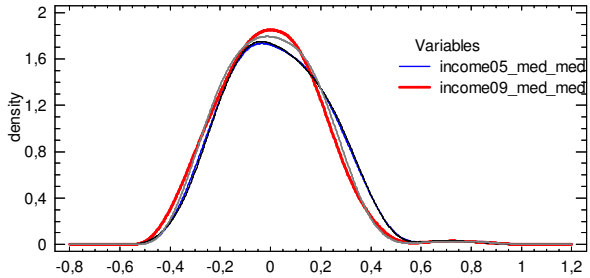
Klesajúce hodnoty kvartilovej šikmosti v období rokov 2007 až 2009 až do záporných hodnôt jednoznačne potvrdzujú náš názor o postupnom narastajúcom posune stredne zarábajúcej vrstvy zamestnancov k nízkym hodnotám príjmov v posledných rokoch. Veľkosť proporcionálneho poklesu/nárastu hornej/dolnej časti v box-plotoch kvartilov, alebo sixtilov je meraná pravou/ľavou šikmosťou. Získané hodnoty potvrdzujú predchádzajúce závery a kvantifikujú veľkosť tohto posunu.



Obrázok 3: Analytický graf priemerných úrovní strednej časti príjmového rozdelenia u zamestnancov v SR v rokoch 2005 až 2009 (s 95 % intervalmi spoľahlivosti)

Výšky priemerných príjmov prostrednej polovice rozdelení v jednotlivých sledovaných rokoch a ich postavenie vzhľadom na ich priemernú hodnotu za celé sledované obdobie je znázornené na obrázku 3. Hranice 95 % - ných intervalov spoľahlivosti potvrdzujú štatisticky významný pokles priemernej úrovne stredne vysokých príjmov na Slovensku od roku 2007.

Našou snahou bolo nájsť primerané štatistické miery, t. j. relatívne a adekvátne pre volené časti príjmového rozdelenia, ako aj názorné grafické zobrazenia vhodne charakterizujúce posun strednej vrstvy zamestnancov k nižším príjmom.



Obrázok 4: Kernelové odhady pdf stredných častí rozdelení relatívnych odchýlok príjmov od mediánov (zamestnanci v SR v rokoch 2005 a 2009 podľa EU-SILC)

Jedným z takýchto analytických grafických zobrazení je porovnanie kernelových odhadov hustôt relatívnych vzdialeností hodnôt v prostredných poloviciach príjmových rozdelení od mediánov príjmov v porovnávaných rokoch. Na osi *x* grafu obrázku 4 sú odchýlky príjmov od mediánu v desatinnom vyjadrení z veľkosti mediánu v príslušnom roku. Prepad empirickej *pdf* v časti nad hodnotami 0,2 až 0,6 v roku 2009 oproti roku 2005 je výrazný a jednoznačne podporujúci naše predchádzajúce tvrdenia.

5. Záver

Výsledky analýz oficiálnych údajov ŠÚ SR v príspevku potvrdzujú tézu o presune stredne zarábajúcej vrstvy zamestnancov v SR v posledných rokoch smerom k nižším príjmom. Zvolili sme relatívne štatistické miery, ktoré vhodne charakterizujú proporcionálny posun v prostrednej polovici príjmových rozdelení k nižším hodnotám, hlavne v období od roku 2007 a tiež grafické zobrazenia, ktoré toto tvrdenie podporili.

Rast príjmovej nerovnosti v SR posunom strednej vrstvy zamestnancov smerom k nižším príjmom v posledných rokoch pri nezmenenom trende znamená jej ohrozenie, ako aj nárast pravdepodobnosti, že sa väčšia časť obyvateľstva SR ocitne bližšie k absolútnej hranici chudoby. Stále menej zamestnancov zo strednej vrstvy má šancu udržať sa na rovnakej úrovni v spoločnosti alebo zaradiť sa k lepšie zarábajúcej vrstve. Táto skutočnosť sa nedá kvantifikovať mierami oficiálne používanými v rámci krajín EÚ, ktoré vychádzajú z *relatívnej* hranice ohrozenia chudobou – stanovenej ako 60 % *národného mediánu* ekvivalentného disponibilného príjmu domácností.

Zhoršujúce sa postavenie strednej vrstvy obyvateľstva, zvlášť v období svetovej hospodárskej recesie, významne negatívne vplýva na agregovaný domáci dopyt a tým aj na potencionálny hospodársky rast a na rast životnej úrovne obyvateľstva.

6. Literatúra

[1]SIPKOVÁ, Ľ. 2010. Porovnanie rozdelení hrubých miezd mužov a žien na Slovensku podľa EU-SILC In: ANALÝZA PRÍJMOVEJ DIFERENCIÁCIE ŽIEN A MUŽOV NA SLOVENSKU : Monografický zborník z riešenia IGP 21/2008 na EU v Bratislave v redakcii Ľubici Sipkovej. Bratislava : EKONÓM, 2010. ISBN 978-80-225-2848-1.

[2]SUTER, Ch. 2010. Globalization, economic and social transformation, and inequality: Introduction to special issue. In: International Journal of Comparative Sociology, August 2010 51, pp. 243-245.

[3]WARD, W. – LELKES, O. – SUTHERLAND, H. – TÓTH, I. T. 2009. European Inequalities: Social Inclusion and Income Distribution in the European Union. Budapest, TÁRKI Social Research Institute Inc., 2009, pp. 214. ISBN 978-963-7869-40-2

Adresy autorov:

Juraj Sipko, doc. Ing. MBA, PhD.
Paneurópska vysoká škola,
Fakulta ekonómie a podnikania
Tematínska 10, 851 05 Bratislava
jsipko@gmail.com

Ľubica Sipková, Ing. PhD.
Ekonomická univerzita v Bratislave
Fakulta hospodárskej informatiky
Dolnozemská 1, Bratislava
lubica.sipkova@euba.sk

Príspevok je súčasťou riešenia výskumného projektu GAČR 402/09/0515.

Kvantilové modely výšky škôd v poistení motorových vozidiel
Quantile Models of Losses in Car Insurance

Ľubica Sipková, Viera Pacáková

Abstract: The objective is to describe the variation in claim amounts by finding a loss distribution that adequately describes the claims which actually occur. Particular attention is paid to the studying of the right tail of the distribution, since it is important to not underestimate the size of large losses. The objective of this paper is to call attention to a new approach to statistical modelling by using inverse distribution functions i. e., quantile functions. The use of models based on quantile methods provides an appropriate and flexible approach to the distributional modelling needed to obtain well-fitted tails. Modern computer simulation techniques open up a wide field of practical applications for this theory concept, without requiring the restrictive assumptions and sophisticated mathematics of many traditional aspects of insurance risk theory.

Key words: Individual Loss Distribution, Quantile Distribution Functions, Gamma-Pareto Probability Model, Lognormal-Pareto Distribution, Öztürk - Dale Estimation Method.

Kľúčové slová: model individuálnych poistných plnení, rozdelenie poistných škôd, kvantilová distribučná funkcia, gamma – Paretov model, lognormálne-Paretovo rozdelenie.

JEL classification: C14, C46, G22

1. Úvod

Účelom aplikácie teórie rizika v poistných vedách je ohodnotiť celkovú výšku poistných plnení. V príspevku venujeme pozornosť jednému z komponentov celkovej sumy poistných plnení – individuálnym poistným plneniam. Celkovú výšku poistných plnení významne ovplyvňujú zriedkavé ale extrémne vysoké poistné plnenia. Aplikácia modelovania poistných škôd pomocou známych pravdepodobnostných funkcií s odhadom ich parametrov metódou maximálnej vierohodnosti často prináša tvary s nedostatočnou zhodou práve v predĺžených pravých koncoch a s podhodnotením extrémnych poistných plnení.

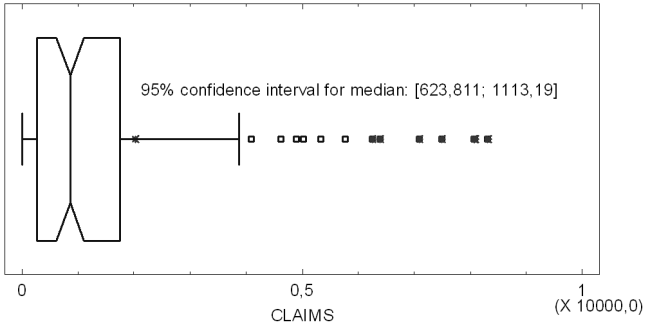
Snahou je aplikovať metódy kvantilového modelovania na rozdelení poistných plnení v neživotnom majetkovom poistení a popísať tak pravdepodobnostné rozdelenie modelom, ktorý prinesie lepšiu zhodu predĺženého pravého konca. Kvalitnejšie modelovanie vysokých hodnôt s primeraným presunutím pravdepodobnosti do konca rozdelenia prináša aj adekvátnejší tvar v jeho strednej časti. Správne, proporcionálne primerané rozdelenie pravdepodobnosti v celom intervale možných hodnôt poistných plnení, znamená lepšiu celkovú zhodu modelu s empirickým rozdelením a v tom zmysle aj kvalitnejší odhad výšok individuálnych poistných plnení aj s možnosťou rýchleho odhadu ich proporcionálneho postavenia v celkovom rozdelení.

Moderné odhady kvantilových tvarov pravdepodobnostných modelov umožňujú aplikácie v poistných vedách bez požiadaviek na splnenie východiskových predpokladov klasických matematicko-štatistických metód pravdepodobnostného modelovania v teórii rizika.

2. Vstupné dáta – poistné plnenia poistenia motorových vozidiel proti krádeži

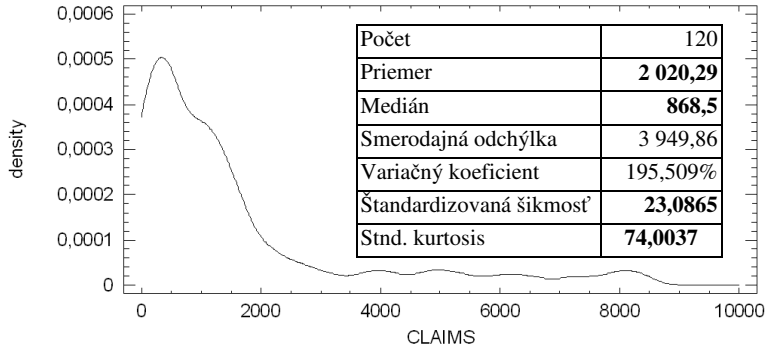
Ako príklad vhodného využitia metód pravdepodobnostného modelovania na kvantilovom základe v prípade poistných plnení v neživotnom poistení slúži aplikácia v príspevku na súbore 120 individuálnych poistných plnení poistenia motorových vozidiel proti krádeži s najnižšou zaznamenanou hodnotou 3 € a najvyššou 32 043 €. Predpokladom je, že tieto individuálne hodnoty poistných plnení sú náhodným výberom z pravdepodobnostného rozdelenia náhodnej premennej *X*, nazývaného rozdelenie poistných škôd (loss distribution).

Grafické zobrazenia vlastností empirického rozdelenia škôd (pozri obrázok 1 a obrázok 2) dopĺňa ich kvantifikácia pomocou základných výberových charakteristík.



Obrázok 2: Box-plot poistných škôd (

Skutočnosť, že rozpätie nad horným kvartilom ($x_{0,75} = 1746$) je približne stokrát väčšie ako rozpätie pod dolným kvartilom ($x_{0,25} = 269$), ako aj počet extrémne vysokých škôd (desať ich prevyšuje hodnotu trojnásobku kvartilového rozpätia pripočítanú k hornému kvartilu) upozorňuje na nesmierne dlhý pravý koniec rozdelenia poistných škôd. (Tri najvyššie škody: 11 453 €; 22 274 €; 32 043 €; nie sú v box-plote zobrazené.)



Obrázok 2: Kernelový odhad funkcie hustoty rozdelenia poistných škôd

Každej vzostupne usporiadanej hodnote individuálneho poistného plnenia, *x*, sme priradili hodnotu pravdepodobnosti *p*, udávajúcu jej proporcionálne postavenie v rade *n* = 120 zistených hodnôt:

Pre $x_{(r:n)}$ odhad pravdepodobnosti je $p_{r:n} = (r - 0{,}5)/n$. Výpočet r-tej hodnoty pravdepodobnosti p zodpovedá rozdeleniu intervalu (0, 1) symetricky do 120 rovnakých častí. Získali sme tak usporiadané dvojice $(x_{(r:n)}, p_{r:n})$, ktoré charakterizujú empirické rozdelenie škôd. Hodnota x, pre každé p sa nazýva výberovým p-kvantilom rozdelenia poistných škôd.

3. Konštrukcia kvantilového modelu výšky poistných škôd

Popri široko využívanej estimácii parametrov funkcií hustôt (*pdfs*) rozdelení poistných škôd (Weibullovho, Paretovho, lognormálneho, gamma rozdelenia) metódou maximálnej vierohodnosti, stále častejšie sa využívajú v pravdepodobnostnom modelovaní aplikácie založené na Galtonom prvýkrát použitých kvantiloch a s metodologickým základom v teórii poriadkových štatistík (Order statistics theory). Pravdepodobnostné modely definované tvarmi inverzných distribučných funkcií (percentilových funkcií) sú známe ako kvantilové pravdepodobnostné modely (QM). Odporúčame ich aplikáciu v pravdepodobnostnom modelovaní poistných škôd v prípade, keď klasický prístup pomocou známych *pdfs* neprináša dostatočnú zhodu horného konca veľmi vysokých škôd.

Hodnoty vierohodnostných pomerov metódou maximálnej vierohodnosti odhadnutých najznámejších tvarov *pdfs* (zosumarizovaných v tabuľke 1) potvrdili najvhodnejší tvar posunutého Weibullovho rozdelenia a posunutého gamma rozdelenia.

Tabuľka 2: Porovnanie odhadnutých pdfs rozdelenia výšky individuálnych škôd

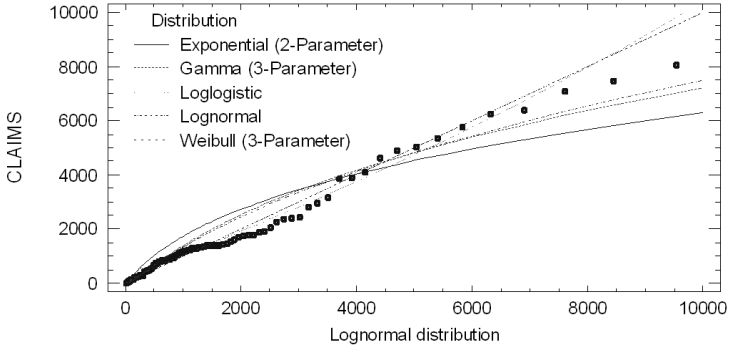
Distribution	*Est. Parameters*	*Log Likelihood*	*Chi-Squared P*	*KS D*	*V*
Weibull (3-Parameter)	3	-1011,67	0,0244651	0,115054	0,154036
Gamma (3-Parameter)	3	-1012,14	0,00878416	0,12502	0,180079
Lognormal (3-Parameter)	3	-1013,12	0,250284	0,0733204	0,12181
Loglogistic	2	-1013,24	0,0556913	0,0674586	0,12815
Loglogistic (3-Parameter)	3	-1013,42	0,331703	0,0739145	0,135035
Lognormal	2	-1014,73	0,1073	0,0867263	0,13881
Weibull	2	-1017,43	0,0219407	0,100633	0,157367
Gamma	2	-1022,46	0,0302371	0,139439	0,18965
Exponential (2-Parameter)	2	-1033,14	0,00141504	0,201559	0,239471
Exponential	1	-1033,32	0,00214622	0,201328	0,239099
Birnbaum-Saunders	2	-1036,11	0,0000017752	0,189219	0,235474
Cauchy	2	-1056,23	$2{,}52777.10^{-9}$	0,217514	0,223384
Largest Extreme Value	2	-1083,48	$9{,}99201.10^{-16}$	0,21277	0,360104
Generalized Logistic	3	-1083,52	$1{,}11022.10^{-16}$	0,213642	0,360952
Laplace	2	-1092,65	0,0	0,296491	0,360887
Exponential Power	3	-1092,65	0,0	0,297527	0,360711
Logistic	2	-1114,82	0,0	0,266996	0,449987
Pareto (2-Parameter)	2	-1120,03	0,0	0,396166	0,609362
Pareto	1	-1141,79	0,0	0,423516	0,654517

Testy dobrej zhody empirického rozdelenia výšky škôd so známymi tvarmi *pdfs* potvrdili vhodnosť aj logaritmicko-normálneho a logistického modelu.

Napriek tomu, podľa Q-Q grafu na obrázku 3, žiadny odhadnutý tvar *pdfs* nemodeluje horný koniec rozdelenia poistných škôd a v dôsledku toho ani jeho strednú časť primerane. (Pozri S-tvar empirického rozdelenia pre logaritmicko-normálny model, ako aj nedostatočné prispôsobenie ostatných tvarov v Q-Q grafe na obrázku 3).

Možno konštatovať, že klasický prístup k pravdepodobnostnému modelovaniu neumožnil nájsť tvar *pdfs*, ktorý by modeloval celé rozdelenie poistných škôd s dobrou zhodou bez dodatočných transformácií empirických údajov. Najlepšiu zhodu priniesli Weibullov, gamma a logaritmicko-normálny tvar, však aj tieto nevhodne modelujú dlhý horný koniec. Okrem možnej matematickej transformácie vstupných hodnôt škôd, riešením je modelovanie prostredníctvom vhodných matematických úprav kvantilových funkcií vyššie

potvrdených vhodných tvarov teoretických rozdelení, prístup známy ako kvantilové pravdepodobnostné modelovanie poistných škôd.



Obrázok 3: Q-Q graf odhadnutých pdfs pre poistné škody

Treba spomenúť aj ďalšie výhody definovania rozdelenia poistných škôd v tvare kvantilovej funkcie (QF), napr. vhodné východisko k Monte-Carlo simuláciám rankitu škôd, ako aj k bootstrapovým odhadom výšky škôd, vysokú elasticitu kvantilových funkcií výhodnú pri modelovaní v podsúboroch a v aktualizáciách modelu poistných škôd v čase.

Možnosť matematických úprav a kombinovania inverzných distribučných funkcií vedie ku konštrukcii primeraného kvantilového modelu. Okrem skladania primeraného kvantilového tvaru treba spomenúť aj možnosť aplikácie niektorého extrémne elastického zovšeobecneného kvantilového tvaru, o čom pojednáva napr. Pacáková, Sipková a Sodomová (2006, s. 263-275). Prvý známy zovšeobecnený tvar lambda rozdelenia (RS GLD), definovaný Rambergom and Schmeiserom (1974, p. 78-82) umožňuje modelovať od záporne asymetrických cez rovnomerné, symetrické, kladne asymetrické tvary s rôznou špicatosťou a dĺžkou ich koncov. Problémom je často presná estimácia jeho štyroch parametrov, z ktorých dva parametre tvaru sú v exponentoch. Metodológiu modelovania pomocou RS GLD možno nájsť napr. v Sipková (2004, p 107-128 or 2005, p. 90-104).

Kvalitnejšie výsledky sme dosiahli konštrukciou koncepčného kvantilového modelu odzrkadľujúceho vlastnosti empirických hodnôt poistných škôd. Opis vhodného matematického aparátu na konštrukciu primeranej kvantilovej funkcie, ako aj techník identifikácie, estimácie a verifikácie kvantilových pravdepodobnostných modelov, zvlášť pre hrubé alebo dlhé konce výrazne asymetrických tvarov rozdelení, poskytuje monografia Sipková a Sodomová (2007) a výklad teórie modelovania kvantilovými tvarmi poskytuje kniha Warrena Gilchrista (2000).

Hlavným cieľom v príspevku je aplikovať metódy koncepčného kvantilového modelovania na celé rozdelenie výšky poistných škôd, posúdiť jeho možnosti aj primeranosť a analyzovať jeho výhody a nevýhody. Preto sa hlbšie nevenujeme výkladu metód kvantilového modelovania, ale len uvádzame odkazy na príslušnú literatúru. Extrémne škody sú vhodne modelované pomocou Paretovho kvantilového tvaru v pravom konci rozdelenia. Modelovaním etrémnych poistných škôd sa zaoberá článok Pacákovej a Sipkovej (2003), alebo Pacákovej, Sipkovej a Šoltésa (2006). Východiskovú teória kvantilového koncepčného modelovania a metód estimácie parametrov QM je časťou teórie štatistickej indukcie pod názvom poriadkové štatistiky Order statistics theory (pozri napr. David and Nagaraja, 2003).

4. Výsledky aplikácie kvantilového modelovania výšky poistných škôd

Použitím pravidiel modifikácie kvantilových funkcií s využitím výsledkov identifikácie najvhodnejších tvarov *pdfs* sme poskladali tri kompozitné kvantilové tvary. Modely boli optimálne vážené tak, aby čo najvhodnejšie rešpektovali vlastnosti vstupných hodnôt škôd.

- Semi-lineárny kvantilový gamma-Paretov model pre poistné škody je:

$$Q(p) = \alpha + \left(\omega (1-p)\, GAMAINV\,(p;\beta.\gamma) + \kappa p \Big/ (1-p)^{\delta} \right),\ 0 < p < 1, \delta > 0 \qquad (2)$$

- Semi-lineárny Weibullov-Paretov kvantilový model je:

$$Q(p) = \lambda + \eta \left\{ (1-p)\left[-\ln(1-p)\right]^{\beta} + p\left[\frac{1}{(1-p)^{\gamma}}\right] \right\},\ 0 < p < 1, \beta > 0, \gamma > 0 \qquad (3)$$

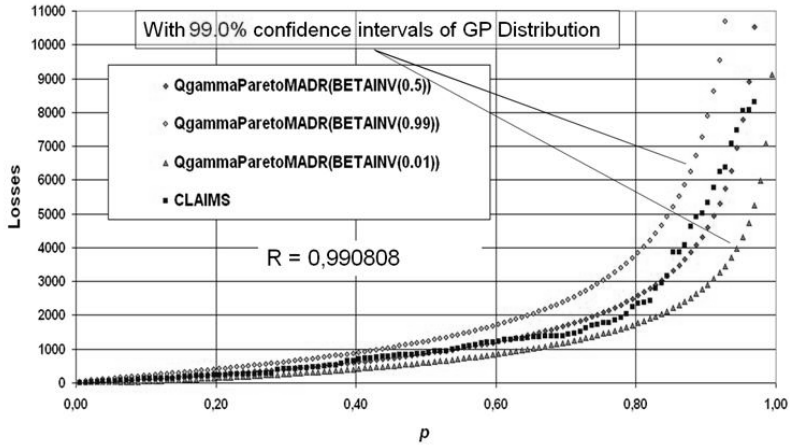
λ je parameter polohy, η je parameter variability, β a γ sú vlastné parametre tvaru.

- Semi-linearny *lognormalny-Paretov kvantilový tvar bez* extra parametra polohy je:

$$Q(p) = \omega (1-p)\, LGINV\,(p;\beta.\gamma) + \kappa p \Big/ (1-p)^{\delta},\ 0 < p < 1, \delta > 0 \qquad (4)$$

V prípade oboch aplikovaných metód estimácie, aproximatívnej zovšeobecnenej Öztürkovej a Daleho metódy a metódy najmenších absolútnych distribučných rezíduí: $e_r = x_{(r)} - Q(p_r;\Theta)$, r = 1, 2, …n. (pozri Gilchrist 2000, s. 198) boli použité kroky minimalizácie optimalizačnej procedúry Solver v Excele.

Najlepšie získané odhadnuté tvary kvantilových modelov rozdelení poistných škôd podľa vzťahov (1 až 3) boli graficky and numericky porovnávané.



Obrázok 3: Porovnanie empirických škôd s gamma-Paretovým modelom pomocou mediánového rankitu a hraníc 99% intervalu spoľahlivosti

Z množstva grafických a numerických metód verifikácie sme uprednostnili tie, ktoré majú priamu súvislosť s kvantilovými charakteristikami. Okrem nich, vhodnou mierou kvality kvantilového modelu je koeficient korelácie medzi simulovanými hodnotami mediánového rankitu odhadnutého kvantilového modelu a empirickými údajmi. V prípade odhadnutých

troch tvarov modelov boli hodnoty koeficientov korelácie vysoké (lognormálny-Paretov model: 0,9898, Weibullov-Paretov model: 0,9905 and gamma-Paretov model: 0,9908). Najlepší odhadnutý tvar aj pre dlhý pravý koniec je gamma-Paretov kvantilový model.

5. Záver

Výsledky jednoznačného kvantilového prístupu v modelovaní rozdelenia poistných škôd v prípade poistenia motorových vozidiel demonštruje jeho účelnosť a využiteľnosť v prípade pravdepodobnostného modelovania individuálnych poistných plnení. Prezentované metódy a techniky prinášajú vhodné tvary pravdepodobnostných modelov a sú aplikovateľné hlavne v prípade komplexných asymetrických rozdelení poistných plnení s existenciou extrémne vysokých zriedkavých hodnôt. Poskladané koncepčné kvantilové tvary modelujú extrémy veľmi vhodne. Naviac, následné analýzy extrémov môžu nadväzovať priamo pri použití kvantilového tvaru v Monte Carlo simuláciách.

6. Literatúra

[12] DAVID, H.A. and NAGARAJA, H.N. 2003. Order Statistics. 3rd Edn., John Wiley and Sons, USA., ISBN: 0-471-38926-9.

[13] GILCHRIST, W.G. 2000, Statistical modelling with quantile functions, Chapman & Hall, ISBN 1-5848-8174-7.

[14] PACÁKOVÁ, V., SIPKOVÁ, Ľ. 2003, Modelovanie extrémnych škôd pomocou kvantilových funkcií, Zborník príspevkov 4. vedeckého seminára Poistná matematika v teórii a v praxi, Bratislava, ISBN 80-225-1771-2.

[15] PACÁKOVÁ, V., SIPKOVÁ, Ľ. 2007, Generalized lambda distributions of household´s incomes. In: E + M. Ekonomie a management : vědecký ekonomický časopis, Roč. 10, č. 1, Technická Univerzita v Liberci, ČR, ISSN 1212-3609.

[16] PACÁKOVÁ, V., SIPKOVÁ, Ľ., SODOMOVÁ, E. 2006, Modelling with Generalized Lambda Distributions. Materiały z XXVII. Ogólnopolskiego Seminarium Naukowego Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych, zorganizovanego przez Zakład Teorii Prognoz Katedry Statystyki Akademii Ekonomicznej w Krakowie (Zakopane, 26-29 apríl 2005); Kraków, ISBN 83-7252-306-1.

[17] PACÁKOVÁ, V., SIPKOVÁ, Ľ., ŠOLTÉS, E. 2006, Simulácia extrémnych škôd pomocou Paretovho rozdelenia. In: Forum Statisticum Slovacum 3/2006, Bratislava: SSDS, SR.

[18] RAMBERG, J., SCHMEISER, B. 1974, An approximate method for generating asymmetric random variables, Communications of the ACM, 17(2), ISSN 0001-0782.

[19] SIPKOVÁ, Ľ. 2004, Zovšeobecnené lambda rozdelenie a odhad jeho parametrov, In: Ekonomika a informatika, 1/2004, Bratislava, SR, ISSN 1336-3514.

[20] SIPKOVÁ, Ľ. 2005, Modelovanie príjmov domácností zovšeobecneným lambda rozdelením. In: Ekonomika a informatika, 1/2005, Bratislava, SR, ISSN 1336-3514.

[21] SIPKOVÁ, Ľ. – SODOMOVÁ, E. 2007, Modelovanie Kvantilovými funkciami, Vydavateľstvo Ekonóm, Bratislava, SR, ISBN 978-80-225-2346-2.

Adresa autorov:

Ľubica Sipková, Ing., PhD.
Ekonomická univerzita, FHI, KŠ
Dolnozemská 1, 852 35 Bratislava
lubica.sipkova@euba.sk

Viera Pacáková, prof., RNDr., PhD.
Univerzita Pardubice
Studentská 95, 532 10 Pardubice 2
vpacakova@gmail.com

Príspevok je súčasťou riešenia výskumného projektu GAČR 402/09/1866 Modelování, simulace a řízení pojistných rizik.

Výpočet variability v heterogenním a homogenním portfoliu
Calculate the variability in Heterogeneous and Heterogeneous Portfolio

Ondřej Slavíček

Abstract: The article deals with the derivation of basic numerical characteristics of homogeneous and heterogeneous portfolio of insurance contracts, if the parameter λ, describing the number of claims during the reporting period, is a random variable with known probability distributions.

Key words: Heterogeneous Portfolio, Homogeneous Portfolio, Variability, Collective Risk, Compound Poisson Distribution

Kľúčové slová: heterogenní portfolio, homogenní portfolio, variabilita, kolektivní riziko, složené Poissonovo rozdělení

JEL classification: C10

1. Úvod

V článku se budeme věnovat rozdílům ve výpočtu variability u heterogenního a homogenního portfolia v závislosti na parametru λ Poissonova rozdělení počtu škod během sledovaného období, který není znám s jistotou, ale jde o náhodnou veličinu, u které známe rozdělení pravděpodobnosti. S Poissonovým rozdělení počtu pojistných událostí pracujeme pro jeho jednoduchost a názornost. Přístupy a myšlenky však lze zobecnit na různá jiná rozdělení, kterými se počet pojistných událostí řídí, například binomické rozdělení nebo negativně binomické rozdělení pravděpodobnosti.

2. Model kolektivního rizika

Dříve než budeme definovat celkovou výši škody, která vznikla během sledovaného období, zaveďme následující značení. Nechť náhodná veličina N udává počet pojistných plnění během sledovaného období a náhodná veličina X_i udává výši i-tého pojistného plnění. Předpokládejme, že náhodné veličiny $\{X_i\}_{i=1}^{\infty}$ jsou nezávislé a mají stejné rozdělení pravděpodobností, u kterého známe první i druhý moment (označme m_1, m_2). Dále předpokládejme, že náhodná veličina N je nezávislá na náhodných veličinách $\{X_i\}_{i=1}^{\infty}$. To jinými slovy znamená, že počet pojistných událostí není závislý na výši jednotlivých pojistných plnění a jednotlivé výše pojistných plnění se vzájemně neovlivňují. Dále předpokládejme, že rozdělení pravděpodobnosti náhodných veličin $\{X_i\}_{i=1}^{\infty}$ se během sledovaného období nemění. Za těchto předpokladů můžeme definovat náhodnou veličinu S určující celkovou výši škody jako:

$$S = X_1 + X_2 + \cdots + X_N.$$

Jestliže náhodná veličina $N = 0$ potom i $S = 0$.

Nyní odvodíme střední hodnotu a disperzi náhodné veličiny S. Protože však S závisí na náhodné veličině N, musíme pro výpočet použít známé vlastnosti podmíněné střední hodnoty a podmíněné disperze. Pro každé dvě náhodné veličiny X, Y platí:

$$E(X) = E\big(E(X|Y)\big) \tag{1}$$

$$D(X) = E\big(D(X|Y)\big) + D\big(E(X|Y)\big). \tag{2}$$

Při hledání E(S) využijeme vlastnost (1) takto:

$$\mathrm{E(S)} = \mathrm{E}\big(\mathrm{E(S|N)}\big), \tag{3}$$

Protože N je náhodná veličina, předpokládejme nyní, že N = n pak

$$\mathrm{E(S|N = n)} = \sum_{i=1}^{n} \mathrm{E(X_i)} = \mathrm{nm_1}, \text{ tedy } \mathrm{E(S|N)} = \mathrm{Nm_1}. \tag{4}$$

Po dosazení do (3) tedy dostáváme:

$$\mathrm{E(S)} = \mathrm{E}\big(\mathrm{E(S|N)}\big) = \mathrm{E(Nm_1)} = \mathrm{E(N)m_1}. \tag{5}$$

Nyní s využitím vlastnost (2) odvoďme disperzi náhodné veličiny S:

$$\mathrm{D(S)} = \mathrm{E}\big(\mathrm{D(S|N)}\big) + \mathrm{D}\big(\mathrm{E(S|N)}\big). \tag{6}$$

Protože náhodné veličiny Xi jsou dle předpokladu nezávislé, platí:

$$\mathrm{D(S|N = n)} = \sum_{i=1}^{n} \mathrm{D(X_i)} = \mathrm{n(m_2 - m_1^2)}, \text{ a tedy } \mathrm{D(S|N)} = \mathrm{N(m_2 - m_1^2)}, \tag{7}$$

po dosazení do (6) dostaneme:

$$\mathrm{D(S)} = \mathrm{E(N(m_2 - m_1^2))} + \mathrm{D(Nm_1)} = \mathrm{E(N)(m_2 - m_1^2)} + \mathrm{D(N)m_1^2}. \tag{8}$$

Podařilo se nám tedy odvodit střední hodnotu a disperzi pro náhodnou veličinu popisující celkovou výši škody. Z výše uvedených vztahů je vidět, že tyto číselné charakteristiky závisí na momentech náhodné veličiny Xi a rozdělení pravděpodobnosti náhodné veličiny N.

V článku dále budeme pracovat s náhodnou veličinou N, která bude mít Poissonovo rozdělení pravděpodobnosti s parametrem λ, tedy $\mathrm{N} \sim \mathrm{Po}(\lambda)$. V takovém případě bude mít náhodná veličina S složené Poissonovo rozdělení s parametrem λ. Střední hodnotu a disperzi náhodné veličiny S určíme pomocí vztahu (5) a (8) a předpokladu:

$$\mathrm{E(N)} = \mathrm{D(N)} = \lambda$$

následovně:

$$\mathrm{E(S)} = \mathrm{E(N)m_1} = \lambda \mathrm{m_1} \tag{9}$$

$$\mathrm{D(S)} = \mathrm{E(N)(m_2 - m_1^2)} + \mathrm{D(N)m_1^2} = \lambda(\mathrm{m_2 - m_1^2}) + \lambda \mathrm{m_1^2} = \lambda \mathrm{m_2}. \tag{10}$$

3. Variabilita v heterogenním portfoliu

Vezměme nyní v úvahu portfolio obsahující *n* nezávislých pojistných smluv. Celkovou výši škody pro *i*-tou pojistnou smlouvu označme jako náhodnou veličinu S_i, kde S_i má složené Poissonovo rozdělení pravděpodobnosti s parametrem λ_i a distribuční funkci F(x). Dále pro jednoduchost předpokládejme, že rozdělení pravděpodobnosti individuální výše škod je stejné pro všechny pojistné smlouvy. Dále předpokládejme, že toto rozdělení je známé na rozdíl od hodnot parametru Poissonova rozdělení λ_i. V této části článku budeme dále předpokládat, že λ_i jsou náhodné veličiny se stejným, známým, rozdělením pravděpodobnosti. To tedy znamená, že jestliže náhodně vybereme pojistnou smlouvu z portfolia, tak předpokládáme, že neznáme příslušný parametr Poissonova rozdělení.

Nyní předpokládejme, že parametr Poissonova rozdělení pojistných smluv v portfoliu není znám, ale se stejnou pravděpodobností se rovná buď hodnotě 0,1 nebo hodnotě 0,3. Naším úkolem je najít střední hodnotu a disperzi celkové výše škody u náhodně vybrané pojistné smlouvy z portfolia a dále pak střední hodnotu a disperzi celkové výše škod celého portfolia.

Z předchozích úvah víme, že $\{\lambda_i\}_{i=1}^n$ jsou nezávislé, stejně rozdělené náhodné veličiny s následujícím rozdělením pravděpodobnosti:

$$P(\lambda_i = 0{,}1) = 0{,}5$$
$$P(\lambda_i = 0{,}3) = 0{,}5$$

a tedy $E(\lambda_i) = 0{,}2$ a $D(\lambda_i) = 0{,}01$.

Nyní vypočítáme střední hodnotu a disperzi náhodné veličiny S_i za podmínky λ_i. Jak jsme již dříve ukázali S_i má složené Poissonovo rozdělení pravděpodobnosti s parametrem λ_i a na základě vztahů (9) a (10) víme, že $E(S_i) = \lambda_i m_1$ a $D(S_i) = \lambda_i m_2$. S využitím těchto vlastností a vztahů (1) a (2) můžeme střední hodnotu a disperzi náhodné veličiny S_i za podmínky λ_i určit jako:

$$E(S_i) = E\big(E(S_i|\lambda_i)\big) = E(\lambda_i m_1) = 0{,}2m_1 \tag{11}$$

$$D(S_i) = E\big(D(S_i|\lambda_i)\big) + D\big(E(S_i|\lambda_i)\big) = E(\lambda_i m_2) + D(\lambda_i m_1) = 0{,}2m_2 + 0{,}01m_1^2. \tag{12}$$

Vypočítali jsme střední hodnotu a disperzi celkové výše škody u náhodně vybrané pojistné smlouvy z portfolia. Nyní se zaměřme na výpočet střední hodnoty a disperze výše škody v celém portfoliu. Protože náhodné veličiny $\{S_i\}_{i=1}^n$ jsou nezávislé a stejně rozložené, můžeme pro výpočet použít výsledky (11) a (12), tedy:

$$E\left(\sum_{i=1}^{n} S_i\right) = nE(S_i) = 0{,}2nm_1 \tag{13}$$

$$D\left(\sum_{i=1}^{n} S_i\right) = nD(S_i) = 0{,}2nm_2 + 0{,}01nm_1^2. \tag{14}$$

4. Variabilita v homogenním portfoliu

Nyní se podívejme na situaci, kdy máme portfolio pojistných smluv homogenní. Opět uvažujme n pojistných smluv, celková výše škody pro jednu pojistnou smlouvu má složené Poissonovo rozdělení pravděpodobnosti s parametrem λ. Tento parametr je na rozdíl od předchozí situace stejný pro všechny pojistné smlouvy. Stejně jako v předchozí situaci předpokládejme, že neznáme hodnotu λ, ale známe rozdělení pravděpodobnosti pro možné hodnoty λ, tedy λ budeme považovat za náhodnou veličinu, a že známe s jistotou první a druhý moment náhodné veličiny popisující individuální výši škody.

Našim úkolem je opět odhadnout střední hodnotu a disperzi celkové výše škody u náhodně vybrané pojistné smlouvy z portfolia a poté také celého portfolia. Náhodná veličina λ má následující rozdělení pravděpodobnosti:

$$P(\lambda = 0{,}1) = 0{,}5$$
$$P(\lambda = 0{,}3) = 0{,}5.$$

S využitím vztahu (1) a (2) můžeme psát:

$$E(S_i) = E\big(E(S_i|\lambda)\big) = E(\lambda m_1) = 0{,}2m_1 \tag{15}$$

$$D(S_i) = E\big(D(S_i|\lambda)\big) + D\big(E(S_i|\lambda)\big) = E(\lambda m_2) + D(\lambda m_1) = 0{,}2m_2 + 0{,}01m_1^2. \tag{16}$$

Vidíme, že střední hodnota i disperze celkové výše škody u náhodně vybrané pojistné smlouvy nám vyšla stejně jako v případě heterogenního portfolia. Nyní se podívejme na výpočet těchto charakteristik u celého portfolia:

$$E\left(\sum_{i=1}^{n} S_i\right) = nE(S_i) = 0{,}2nm_1 \tag{17}$$

$$D\left(\sum_{i=1}^{n} S_i\right) = E\left(D\left(\sum_{i=1}^{n}(S_i|\lambda)\right)\right) + D\left(E\sum_{i=1}^{n}(S_i|\lambda)\right) = E(n\lambda m_2) + D(n\lambda m_1) = 0{,}2nm_2 + 0{,}01n^2m_1^2. \tag{18}$$

Při porovnání výsledků (13) a (17) vidíme, že střední hodnota celkové výše škody u celého portfolia je stejná pro homogenní i heterogenní portfolio. Když ale porovnáme vztahy (14) a (18), vidíme, že disperze celkové výše škody je větší pro homogenní portfolio.

5. Závěr

V článku jsou odvozeny základní číselné charakteristiky, jako je střední hodnota a disperze, náhodné veličiny celková výše škody jak pro náhodně vybranou pojistnou smlouvu z portfolia, tak pro celé portfolio smluv v závislosti na hodnotách λ popisujících počet pojistných událostí během sledovaného období. Uvažujeme přitom možnost, že hodnota λ není dána s jistotou, ale jedná se o náhodnou veličinu se známým rozdělením pravděpodobnosti. V článku je pro jednoduchost uvažované alternativní rozdělení parametru λ. Výše popsané číselné charakteristiky byly odvozeny pro homogenní i heterogenní portfolio.

6. Literatura

[1] DICKSON, D. – WATERS, H. R. *Risk models*. Edinburgh, 1993. 67 s.
[2] PACÁKOVÁ, V. *Aplikovaná poistná štatistika.* 3., preprac. a dopl. vyd. Bratislava : Iura Edition, 2004. 261 s. ISBN 80-8078-004-8(brož.).

Jméno autora

Ondřej Slavíček, Mgr.
Ústav matematiky, Fakulta ekonomicko-správní
Univerzita Pardubice,
Ondrej.Slavicek@upce.cz

Tento článek vznikl jako součást řešení projektu GAČR 402/09/1866: Modelování, simulace a řízení pojistných rizik.

Využitie dataminingových nástrojov pri hľadaní nových liečiv
Using of data mining tools in drug design

Mária Stachová, Miroslav Medveď

Abstract: The central task of this paper is to present a possible mathematical classification models that are useful to predict the biological activity of a molecule in a high accuracy and thus can help us to identify the candidates for new drugs.

The Classification Tree, The Conditional Classification Tree, The Random Forest and The Conditional Random Forest classification algorithms are used to denote molecular descriptors that share to molecular biological activity. In these tree structures, leaves represent classifications and branches represent conjunctions of features that lead to these classifications.

Key words: The Decision Tree, The Conditional Decision Tree, The Random Forest, The Conditional Random Forest, data mining.

Kľúčové slová: klasifikačný strom, podmienený klasifikačný strom, náhodný les, podmienený náhodný les, data mining.

JEL classification: C19

1. Úvod

V literatúre môžeme nájsť viacero definícií pojmu datamining. Fayyad definuje pojem nasledovne: Datamining (DM) je netriviálny proces zisťovania platných, neznámych, potenciálne užitočných a ľahko pochopiteľných závislostí v dátach [4]. Jedná sa teda o "dolovanie z dát". Môžeme povedať, že datamining je všeobecná metodológia , ktorá sa používa na riešenie rôznych problémov, sledovaním toku dát, monitorovaním procesu s predikciou vývoja či zlyhania, odhadom budúceho správania sa individuálnych prípadov a pod.

Medzi často používané metódy DM patria okrem iných aj klasifikačné stormy, náhodné lesy, modifikácie klasifikačných stromov a náhodných lesov, ale aj klasické štatistické techniky, ako logistická regresia. V našej práci venujeme pozornosť práve týmto technikám. Prvé vymenované metódy sme vybrali preto, lebo sa ukazujú ako dostatočne presné prediktory v oblasti našej aplikácie, t.j. v oblasti molekulovej chémie, pričom logistickú regresiu sme vybrali skôr na porovnanie klasických štatistických a DM nástrojov.

2. Dátová množina

Našou snahou je nájsť vhodný dataminingový model, ktorý by s čo najväčšou presnosťou predpovedal vzťah medzi štruktúrou molekúl a ich biologickou aktivitou.

Biologická aktivita molekúl závisí od niektorých ich chemických a fyzikálnych vlastností (molekulových deskriptorov) a popisuje ju inhibičný koeficient $IC50$, nazývaný aj inhibičná koncetrácia. $IC50$ udáva koncentráciu danej látky, ktorá spôsobí 50%-né spomalenie skúmanej biochemickej reakcie. Čím nižšia je jeho hodnota, tým vyššia je biologická aktivita danej molekuly.

Molekulové deskriptory sú fyzikálne a chemické vlastnosti molekúl, vyjadrené číslami, ktoré opisujú ich štruktúru a tvar. Pomáhajú predpovedať aktivitu a vlastnosti samotnej molekuly. Sú odvodené z ich chemickej štruktúry, konštitúcie, väzbovosti, 3D geometrie, rozloženia náboja v molekule a pod. Príkladmi molekulových deskriptorov sú okrem

mnohých iných aj molekulová hmotnosť, dipólový moment, počet aromatických jadier v molekule, refraktivita a počet vodíkových väzieb, ktoré je molekula ochotná odovzdať, alebo prijať. Molekulové deskriptory nám teda slúžia ako prediktory, na základe ktorých sa snažíme predpovedať ako sa daná molekula bude správať v biochemických reakciách.

Zaujímal nás konkrétny druh biologickej aktivity týchto molekúl a tým je ich VEGFR-2 (Vascular Endothelial Growth Factor Receptor-2) inhibičná koncentrácia *IC*50, potrebná na zastavenie novotvorby ciev (angiogenéza, neovaskularizácia) zásobujúcich nádor. Nádor trpí nedostatkom a nemôže ďalej rásť. Inhibítory angiogenézy sú teda vhodné liečivá, využiteľné pri spomalení, či zastavení rastu tumorov a ich metastáz. Ich toxicita, je oproti bežným chemoterapeutikám veľmi nízka, pričom mechanizmus pôsobenia týchto látok na tumor je odlišný od klasických chemoterapeutík [1].

Pre našu analýzu sme použili databázu 1282 molekúl (trénovacia množina), so známou biologickou aktivitou *IC*50 (závislá premenná). Pomocou chemického softvéru JChem 5.1.2 sme pre tieto molekuly vypočítali 32 deskriptorov, ktoré sme si označili D1-D32 a na základe experimentálnych vedomostí sme si určili 3 rôzne hranice závislej premennej a to *IC*50 = 50, *IC*50 = 150 a *IC*50 = 500. Molekuly, ktorých *IC*50 je menšie ako prvá hranica sa dajú považovať za veľmi dobre biologicky aktívne a naopak molekuly, ktorých *IC*50 je vyššie ako tretia hranica sa môžu považovať za biologicky pasívne. Pre každú hranicu *IC*50 sme vybudovali klasifikačné štruktúry popísané v nasledujúcej časti.

3. Metodológia

Kvôli mnohým výhodám ako sú napríklad názornosť, možnosť extrahovať pravidlá, ktoré sú na pozadí modelu, ako aj menšia náročnosť na prípravu dát (napr. nepožadujeme normálne rozdelenie a pod.) sme sa v našej analýze zamerali na klasifikačné stromy a podmienené klasifikačné stromy. Tieto metódy porovnávame navzájom z pohľadu ich schopnosti predpovedať hodnoty závislej premennej, ktorou je u nás biologická aktivita molekúl pri rôznej hladine inhibičného koeficientu *IC*50. Tak isto ich porovnávame aj s klasifikátormi, ktoré sú na týchto „stromových“ štruktúrach založené a to s rozhodovacími stromami a podmienenými rozhodovacími stromami. Nakoniec všetky výsledky sú porovnané s výsledkami získanými logistickou regresiou.

Klasifikačné stromy (Classification trees, alebo tiež rozhodovacie stromy, Decision trees) sú jednou z metód DM používaných na klasifikáciu dát a predikciu. Je to štruktúra vybudovaná pomocou trénovacej databázy, ktorá na základe vstupných dát (prediktorov) predpovedá hodnotu výstupnej premennej. My sme použili tzv. CART algoritmus (Classification and Regression Tree), popísaný v [3], ktorý je implementovaný do programu *R*, pomocou knižnice *rpart* [7].

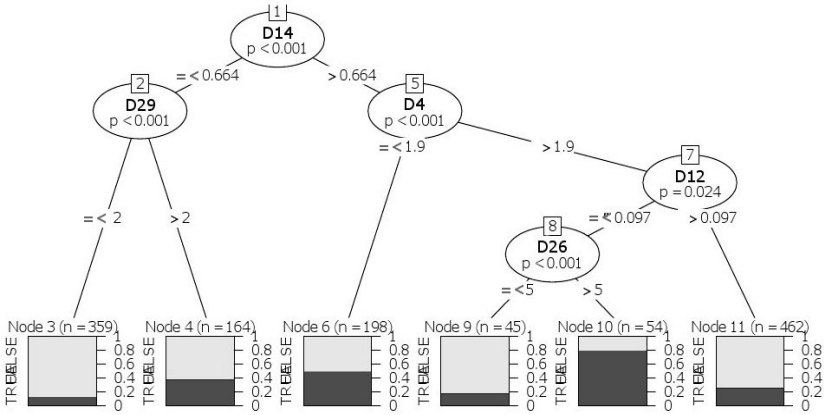
Model **podmienených klasifikačných stromov** (cTree) odhaduje regresný vzťah pomocou rekurzívneho delenia dát v podmienenej inferenčnej štruktúre [5]. V tomto algoritme sa testuje celkov nulová hypotéza nezávislosti medzi všetkými vstupnými premennými a nezávislými premennými. Ak sa hypotéza nezamieta algoritmus skončí, ak sa ale nulová hypotéza zamieta, algoritmus vyberie vstupnú premennú s najsilnejšou asociáciou na závislú premennú. V ďalšom kroku je binárne delenie realizované na vybranej vstupnej premennej. Tieto kroky sa rekurzívne opakujú. Kritériom pre ukončenie algoritmu je najmenšia hladina testu (p-hodnota), ktorej by sme ešte danú hypotézu zamietli. Aj tento model je implementovaný v programe *R*, vyvolaný funkciou cTree pomocou knižnice *party* [5].

Na Obrázku 1 je príklad vytvoreného podmieneného rozhodovacieho stromu, ktorý bol vybudovaný pomocou programu R (knižnica *party*) pre hladinu *IC*50=50. Strom pozostáva z tzv. rodičovského uzla, v ktorom sú obsiahnuté všetky molekuly. Na základe hodnoty

prediktora označeného ako D14 sa tento uzol rozdeľuje do dvoch vetiev. Molekuly, ktorých je hodnota tohto prediktora menšia, alebo rovná 0,664 prechádzajú do uzla o úroveň nižšie, kde sa zas delia na základe ďalšieho deskriptora označeného ako D29 do koncových uzlov. Podobne sa klasifikujú aj molekuly, ktorých hodnota deskriptora D14 bola väčšia ako 0,664.

Náhodný les je model pozostávajúci z mnohých klasifikačných stromov, z ktorých sa vytvorí „volebná komisia" a o výsledku (teda u nás o zaradení molekuly do príslušnej triedy) sa rozhodne „hlasovaním". Autorom myšlienky náhodného lesa je Leo Breiman a bližšie je popísaná v [2]. Aj pri jeho budovaní sme využili štatistický program *R* a tentoraz knižnicu *randomForest*.

Model **podmienených náhodných lesov** je algoritmus tvorený klasifikačnými stromami v podmienenej inferenčnej štruktúre [5]. Tento model je vylepšením cTree algoritmu, pričom je často robustejší. Nevýhodou je, že jeho výstup nie je tak jednoducho interpretovateľný ako to bolo pri stromovom modele. V našej práci sme tento klasifikátor vybudovali pomocou *party* knižnice v programe *R*.



Obrázok 3: Podmienený klasifikačný strom zostrojený pre hladinu IC50=50

4. Výsledky

Pri tvorbe klasifikačných modelov sme si najskôr na základe experimentálnych vedomostí určili hranice parametra IC50 (v hodnotách IC50 = 50, IC50 = 150 a IC50 = 500) a na základe nich sme potom v štatistickom programe *R* pomocou knižníc *rpart*, *randomForest* a *party* postupne vybudovali žiadané modely.

Kvalitu jednotlivých modelov popisujú ich tabuľky (matice) úspešnosti, z ktorých môžeme vyčítať schopnosť modelu predpovedať. Riadky v maticiach predstavujú skutočné hodnoty zaradenia objektov a maticové stĺpce predstavujú predikované hodnoty. Diagonálne hodnoty patria správne zaradeným objektom a krížové diagonálne hodnoty patria nesprávne zaradeným objektom. V Tabuľke 1 sú tabuľky úspešnosti pre naše vybudované modely. Prvý riadok patrí modelom zostrojených pre hranicu IC50 = 50, druhý a tretí patrí modelom zostrojených pre hranicu IC50 = 150 a IC50 = 500. Pre všetky modely bola vypočítaná aj chybovosť a taktiež je uvedená v Tabuľke 1. V poslednom stĺpci tabuľky sú matice úspešnosti logisticej regresie, tak isto uvádzané s jej chybovosťami.

5. Záver

Výsledky uvedené v Tabuľke 1 naznačujú, že použité dataminingové nástroje (CART algoritmus, podmienený klasifikačný strom, podmienený klasifikačný strom, náhodný les, podmienený náhodný les a logistická regresia) sa ukázali byť efektívne pri hľadaní vzťahu medzi štruktúrou a vlastnosťami (biologickou aktivitou) pre náš trénovací súbor objektov, ktorý tvorili molekuly, potenciálne liečivá. Podľa najnižšej chybovosti môžeme usudzovať, že ako veľmi dobrý klasifikátor sa javí podmienený náhodný les. Užitočné sú ale aj samostatné stromové štruktúry (klasifikačné stromy a podmienené klasifikačné stromy), pretože sú názorné a dokážu triediť databázu neznámych objektov postupným delením podľa jednotlivých prediktorov.

Tabuľka 3: Tabuľky úspešnosti jednotlivých modelov

prah IC_50/logIC_50	Klasifikačný strom	Podmienený klasifikačný strom	Náhodný les	Podmienený náhodný les	Logistická regresia
50 / 1.699	False True False 850 57 True 207 168 error rate: **20.59%**	False True False 896 11 True 332 43 error rate: **26.76%**	False True False 835 72 True 188 187 error rate: **20.28%**	False True False 823 84 True 185 190 error rate: **20.98%**	False True False 852 55 True 295 80 error rate: **27.30%**
150 / 2.176	False True False 531 141 True 253 357 error rate: **30.73%**	False True False 599 73 True 439 171 error rate: **39.94%**	False True False 515 157 True 176 434 error rate: **25.98%**	False True False 516 156 True 174 436 error rate: **25.74%**	False True False 449 223 True 277 333 error rate: **26.21%**
500 / 2.699	False True False 199 194 True 133 756 error rate: **25.51%**	False True False 39 354 True 26 863 error rate: **29.64%**	False True False 189 204 True 91 798 error rate: **23.01%**	False True False 194 199 True 94 795 error rate: **22.85%**	False True False 120 273 True 57 832 error rate: **23.48%**

6. Literatúra

[1] ADDOVÁ G. 2010. Vývoj nových neo-plastických liečív. In. Posterus.sk, ISSN 1338-0087, marec (2010), http://www.posterus.sk/?p=5009.

[2] BREIMAN L. 2001. Random forests. Machine Learning. 45, 2001, s. 5 – 32.

[3] BREIMAN L. - FRIEDMAN J.H. - OLSHEN R.A. - STONE C.J. 1984. Classification and regression trees. Belmont CA: Wadsworth.

[4] FAYYAD U. – PIATETSKY-SHAPIRO G. – SMYTH P. 1996, Proceeding of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon.

[5] HOTHORN T. - HORNIK K. - ZEILIS A. 2006. Unbiased recursive partitioning: a conditional inference framework. In: Journal of Computational and Graphical Statistics. 15(3), 2006, s. 651-674.

[6] HOTHORN - T., ZEILIS A. - HORNIK K. 2007. Let's have a party! an open-source toolbox for recursive partytioning. Research Report Series 59. Department of Statistics and Mathematics, Wirtschaft-suniversität Wien, Wien, Austria.

[7] THERNEAU T.M. - ATKINSON, E.J. 1997 . An Introduction to Recursive Partitioning Using the RPART Routines [online], [cit. 25.8.2009].

Adresa autorov:

Mária Stachová, Mgr., PhD.
EF UMB, Tajovského 10, 974 00 Banská Bystrica
maria.stachova@umb.sk

Miroslav Medveď, Doc., RNDr., PhD.
FPV UMB, Tajovského 40, 974 00 Banská Bystrica
miroslav.medved@umb.sk

Príspevok bol napísaný s podporou Univerzitnej grantovej agentúry Univerzity Mateja Bela v Banskej Bystrici v rámci riešenia vedecko-výskumných projektov UGA I-10-005-07 a UGA I-09-00029.

Symetrická reprezentácia polynómov
Symmetric representation of polynomials

Imrich Szabó, Csaba Török

Abstract: The information gain needs qualitative numerical and statistical techniques. For modeling and smoothing noisy 2D data with complex structure we recently proposed a three-part local scheme in which the key role was played by a special representation of polynomials based on four reference points. The goal of this article is to derive a similar representation for polynomials of two variables that should play the central role in modeling and smoothing 3D data.

Keywords: Reference points, interpolation, polynomial representation, surfaces

Kľúčové slová: Referenčné body, interpolácia, polynomiálna reprezentácia, plochy

1. Úvod

Pri súčasnej úrovni informatizácie vo svete predstavujú dáta (či už zozbierané pomocou počítača alebo inou formou) mimoriadne dôležitý element, s ktorým sa stretávame na každom kroku a spod jeho vplyvu sa nemožno vymaniť – človeka dnešnej doby sprevádzajú dáta už od narodenia.

V tomto článku však dostanú priestor hlavne dáta experimentálne, zaťažené chybami – hovorí sa, že dáta sú zašumené. Ako takéto dáta vyhladiť? V prípade dvojrozmerných dát dobre poslúži napríklad klasická regresia. Ak však majú zašumené dáta komplexnejšiu štruktúru, metóda najmenších štvorcov nemusí stačiť (často krát ani nestačí) a výsledky nedopadnú podľa očakávania. Jedna z možností, ako sa s touto situáciou vysporiadať je úseková aproximácia, ktorá spočíva v rozdelení aproximovanej plochy na menšie oblasti zvané segmenty. Na tie sa (podľa určitého vzoru) aplikuje klasická regresia a nová metóda založená na IZA (Interpolation - Zeroing – Approximation polynomial) reprezentácii polynómov a vhodnom rozmiestnení referenčných bodov s cieľom dosiahnut hladký prechod medzi jednotlivými segmentami. Takýmto spôsobom je možné vytvoriť elementárne schémy, ktoré sa budú dať aplikovať na zložitejšie štruktúry.

Obsah tohto článku je rozdelený na dve hlavné časti. Prvá z nich pomáha čitateľovi zorientovať sa v oblasti transformácií a ich vplyvov na mocninové funkcie či polynómy jednej premennej. Druhá, najdôležitejšia časť, sa venuje konštrukcii špeciálnej IZA reprezentácie polynómov, ktorá bude mať dôležitú funkciu pri vyhladzovaní trojrozmerných dát poznačených chybami. Ukáže sa aj spôsob vylepšenia tejto reprezentácie z hľadiska symetrie zápisu či odstránenia duplicitných výskytov koeficientov.

2. Stručne o transformáciách

Táto kapitola okrem definícií doprednej a inverznej transformácie pojednáva aj o ich vplyve na mocninové funkcie a polynómy jednej premennej. Dôvod, prečo sa tu spomínajú, je ich nevyhnutná účasť pri dokazovaní tvrdení ohľadne IZA reprezentácie polynómov. O vplyvoch týchto transformácií na polynómy a funkcie dvoch premenných je možné sa dozvedieť viac z [3].

Definícia 1
Dopredná štvorbodová dvojrozmerná transformácia ľubovoľnej spojitej funkcie jednej premennej *f(x),* založená na množine referenčných bodov $R \equiv R_v = \{[v_i f(v_i)]; i = 0,\ldots,3\}$, je definovaná vzťahom

$$T_{R_v} f(x) = H_{R_v,0}(x) f(x) + \sum_{i=1}^{3} H_{R_v,i}(x) f(v_i),$$

kde $H_{R_v,i}(x)$, pre $i = 0,\ldots,3$ sú dané nasledovne

$$H_{R_v,0}(x) = \frac{(v_0 - v_1)(v_0 - v_2)(v_0 - v_3)}{(x - v_1)(x - v_2)(x - v_3)}, \quad H_{R_v,1}(x) = \frac{(v_0 - x)(v_0 - v_2)(v_0 - v_3)}{(v_1 - x)(v_1 - v_2)(v_1 - v_3)},$$

$$H_{R_v,2}(x) = \frac{(v_0 - x)(v_0 - v_1)(v_0 - v_3)}{(v_2 - x)(v_2 - v_1)(v_2 - v_3)}, \quad H_{R_v,3}(x) = \frac{(v_0 - x)(v_0 - v_1)(v_0 - v_2)}{(v_3 - x)(v_3 - v_1)(v_3 - v_2)}.$$

Definícia 2
Inverznú transformáciu ľubovoľnej spojitej funkcie jednej premennej *f(x,)* založenú na množine štyroch referenčných bodov $R \equiv R_v = \{[v_i f(v_i)]; i = 0,\ldots,3\}$, vyjadrujeme nasledujúcim vzťahom

$$T_{R_v}^{-1} T_{R_v} f(x) = \Pi_{R_v,0}(x) T_{R_v} f(x) + \sum_{i=1}^{3} \Pi_{R_v,i}(x) f(v_i),$$

kde $\Pi_{R_v,i}(x)$, pre $i = 0,\ldots,3$ sú dané nasledovne

$$\Pi_{R_v,0}(x) = \frac{(x - v_1)(x - v_2)(x - v_3)}{(v_0 - v_1)(v_0 - v_2)(v_0 - v_3)}, \quad \Pi_{R_v,1}(x) = \frac{(x - v_0)(x - v_2)(x - v_3)}{(v_1 - v_0)(v_1 - v_2)(v_1 - v_3)},$$

$$\Pi_{R_v,2}(x) = \frac{(x - v_0)(x - v_1)(x - v_3)}{(v_2 - v_0)(v_2 - v_1)(v_2 - v_3)}, \quad \Pi_{R_v,3}(x) = \frac{(x - v_0)(x - v_1)(x - v_2)}{(v_3 - v_0)(v_3 - v_1)(v_3 - v_2)}.$$

Poznámka
Namiesto celej množiny R_v budeme často pracovať len s vektorom prvých súradníc referenčných bodov $v=(v_0,\ldots,v_3)$. Taktiež budeme písať $T_v f(x)$, respektíve $T_{x,v} f(x)$ namiesto $T_{R_v} f(x)$ a $p_i^v(x)$ namiesto $\Pi_{R_v,i}(x)$. To isté platí aj pre inverznú transformáciu.

Uvádzame dve tvrdenia o 2D transformácii mocninových funkcií a polynómov jednej premennej z [3] a [4]

Lema 1 (Štvorbodová 2D transformácia mocninovej funkcie)
Nech je daná mocninová funkcia jednej premennej $f(x) = x^p$, pre $p \in Z^{0+}$. Potom štvorbodová dvojrozmerná transformácia tejto funkcie bude

a) konštanta pre $p < 4$ a platí

$$T_v x^p = v_0^p,$$

b) polynóm stupňa *p-3* pre $p \geq 4$ a platí

$$T_v x^p = v_0^p + z_1^v(x) S_{p-4,4}^v,$$

kde

$$z_1^v(x) = (x - v_0) \prod_{i=1}^{3} (v_0 - v_i) \tag{1}$$

a $S_{p-4,4}^v$ sa počíta na základe vzťahu

$$S_{j,k}^v = S_{j,k-1}^v + v_k S_{j-1,k}^v, \tag{2}$$

pre $1 \leq j$; $1 \leq k \leq 4$, kým $S_{0,k}^v = 1, k \geq 1$, $S_{j,0}^v = v_0^j, j \geq 1$, $S_{1,k}^v = \sum_{i=0}^{k} v_i$, $k \geq 1$ a $v_4 \equiv x$.

Príklad 1
Nech $p = 5$, *resp. 7*. Potom výsledok transformácie funkcie $f(x) = x^p$, použitím vektora $v=(v_0,...,v_3)$ bude

$$T_v x^5 = v_0^5 + (x - v_0)(v_0 - v_1)(v_0 - v_2)(v_0 - v_3)(v_0 + v_1 + v_2 + v_3 + x),$$

respektíve

$$T_v x^7 = v_0^7 + (x - v_0)(v_0 - v_1)(v_0 - v_2)(v_0 - v_3)(v_0^3 + v_1(v_0^2 + v_1(v_0 + v_1)) + v_2(v_0^2 + v_1(v_0 + v_1)$$
$$+ v_2(v_0 + v_1 + v_2)) + v_3(v_0^2 + v_1(v_0 + v_1) + v_2(v_0 + v_1 + v_2) + v_3(v_0 + v_1 + v_2 + v_3)) +$$
$$x.(v_0^2 + v_1(v_0 + v_1) + v_2(v_0 + v_1 + v_2) + v_3(v_0 + v_1 + v_2 + v_3) + x(v_0 + v_1 + v_2 + v_3 + x))).$$

Veta 1 (Štvorbodová 2D transformácia polynómu)

Nech p je zo Z^{0+}. Potom štvorbodová dvojrozmerná transformácia polynómu $P_p(x) = \sum_{i=0}^{p} a_i x^i$,

pri použití vektora $v=(v_0,...,v_3)$ bude

a) konštanta pre $p < 4$ a platí

$$T\, P_p(x) = P_p(v_0),$$

b) polynóm stupňa *p-3* pre $p \geq 4$ a platí

$$T\, P_p(x) = P_p(v_0) + z_1^v(x) A_{p-4}^v(x),$$

kde $z_1^v(x)$ je definované vo vzťahu (1) a

$$A_{p-4}^v(x) = \sum_{i=4}^{p} a_i S_{i-4,4}^v. \tag{3}$$

3. IZA reprezentácia polynómov

Táto kapitola predstavuje hlavnú časť článku zastrešujúcu najdôležitejšie zistenia doterajšieho výskumu, ktoré sa opierajú o výsledky uvedené či už v predchádzajúcej kapitole, alebo v prácach [3] a [4].

Pre dôkaz vety o 4x4 bodovej IZA reprezentácii polynómu $P_{p,q}(x,y)=\sum_{i=0}^{p}\sum_{j=0}^{q}a_{ij}x^i y^j$ budeme okrem štvorbodovej doprednej a inverznej transformácie potrebovať aj nasledujúce pomocné lemy, dôkazy ktorých sú uvedené v [4].

Lema 2
Nech $p,q\in Z$, $p\geq 4$, $q\geq 4$ a $\mu=(\mu_0,...,\mu_3)$. Potom platí nasledujúci vzťah

$$T_{x,\mu}P_{p,q}(x,y)=P_{p,q}(x_0,y)+z_1^{\mu}(x)A_{p-4}^{\mu}(x,y),$$

kde $z_1^{\mu}(x)$ je založené na (1) a $A_{p-4}^{\mu}(x,y)=\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S_{i-4,4}^{\mu}(x)\,y^j$.

Kým táto lema vyjadruje výsledok štvorbodovej transformácie polynómu $P_{p,q}(x,y)$ podľa premennej x, nasledujúca lema naopak vyjadruje výsledok štvorbodovej transformácie polynómu $P_{p,q}(x_k,y)$ podľa premennej y.

Lema 3
Nech $p,q\in Z$, $p\geq 4$, $q\geq 4$ a $\nu=(\nu_0,...,\nu_3)$. Potom platí nasledujúca rovnica

$$T_{y,\mu}P_{p,q}(x_k,y)=P_{p,q}(x_k,y_0)+z_1^{\nu}(y)A_{q-4}^{\nu}(x_k,y),$$

kde $z_1^{\nu}(y)$ sa počíta podľa (1) a $A_{q-4}^{\nu}(x_k,y)=\sum_{j=4}^{q}\sum_{i=0}^{p}a_{ij}\,S_{j-4,4}^{\nu}(y)\,x^i$.

Nasledujúci výsledok o explicitnom vyjadrení polynómu dvoch premenných pomocou šestnástich referenčných bodov bol dokázaný taktiež v [4]. Uvedená práca obsahuje aj ilustračný príklad na objasnenie princípu fungovania tohto vyjadrenia.

Veta 2 (IZA reprezentácia polynómov)
Nech $p,q\geq 4$, $\mu=(\mu_0,...,\mu_3)$ a $\nu=(\nu_0,...,\nu_3)$. Potom sa dá polynóm $P_{p,q}(x,y)=\sum_{i=0}^{p}\sum_{j=0}^{q}a_{ij}x^i y^j$ vyjadriť nasledujúcim spôsobom

$$P_{p,q}(x,y)=I_{3,3}^{\mu,\nu}(x,y)+z_4^{\nu}(y)A_{q-4}^{\nu}(x_k,y)+z_4^{\mu}(x)A_{p-4}^{\mu}(x,y),$$

kde
$$I_{3,3}^{\mu,\nu}(x,y)=\sum_{i=0}^{3}\sum_{j=0}^{3}p_i^{\mu}(x)\,p_j^{\nu}(y)P_{p,q}(x_i,y_j), \tag{4}$$

$z_4^{\mu}(x)$ a $z_4^{\nu}(y)$ sú dané súčinmi $p_0^{\mu}(x)z_1^{\mu}(x)$, respektíve $p_0^{\nu}(y)z_1^{\nu}(y)$. $z_1^{\mu}(x)$ a $z_1^{\nu}(y)$, ako aj $A_{p-4}^{\mu}(x,y)$, $A_{q-4}^{\nu}(x_k,y)$ sú definované v lemách 2 a 3.

Pri bližšom pohľade na vetu 2 si možno všimnúť určitú asymetriu v danej reprezentácii. Pri rozbore vety taktiež zistíme, že pravý dolný roh matice je zahrnutý v oboch komponentoch *A*, teda v $A^{\nu}_{q-4}(x_k, y)$ a $A^{\mu}_{p-4}(x, y)$. Nasledujúca veta ponúka alternatívu v podobe novej reprezentácie a tretej komponenty *A* počítajúcej pravý dolný roh matice, ktorá tým pádom bráni vzniku prekrývania (duplicitného výskytu koeficientov).

Veta 3 (IZA reprezentácia polynómov - revízia)

Nech $p, q \geq 4$, $\mu = (\mu_0, ..., \mu_3)$ a $\nu = (\nu_0, ..., \nu_3)$. Potom sa dá polynóm $P_{p,q}(x, y) = \sum_{i=0}^{p}\sum_{j=0}^{q} a_{ij} x^i y^j$ vyjadriť v nasledujúcom tvare

$$P_{p,q}(x, y) = I^{\mu,\nu}_{3,3}(x, y) + z^{\nu}_4(y) A^{\nu}_{q-4}(x_k, y) + z^{\mu}_4(x) A^{\mu}_{p-4}(x, y_l) + A^{\mu,\nu}_{p-4,q-4}(x, y).$$

$I^{\mu,\nu}_{3,3}(x, y)$ je definovaný v (4),

$$A^{\nu}_{q-4}(x_k, y) = \sum_{i=0}^{3}\sum_{j=4}^{q} a_{ij} S^{\nu}_{j-4.4}(y) \sum_{k=0}^{3} p_k(x).x_k^i, \qquad A^{\mu}_{p-4}(x, y_l) = \sum_{i=4}^{p}\sum_{j=0}^{3} a_{ij} S^{\mu}_{i-4.4}(x) \sum_{l=0}^{3} p_l(y).y_l^j,$$

$$A^{\mu,\nu}_{p-4,q-4}(x, y) = \sum_{i=4}^{p}\sum_{j=4}^{q} a_{ij} \left(z^{\mu}_4(x) S^{\mu}_{i-4.4}(x). \sum_{l=0}^{3} p_l(y).y_l^j + z^{\nu}_4(y) S^{\nu}_{j-4.4}(y). \sum_{k=0}^{3} p_k(x).x_k^i + z^{\mu}_4(x) z^{\nu}_4(y) S^{\mu}_{i-4.4}(x) S^{\nu}_{j-4.4}(y) \right)$$

a $z^{\mu}_4(x), z^{\nu}_4(y)$ sa konštruujú identicky ako v predchádzajúcej vete.

Dôkaz

Pretože $P_{p,q}(x, y) = T^{-1}_{y,\nu} T_{y,\nu} \left(T^{-1}_{x,\mu} T_{x,\mu} P_{p,q}(x, y) \right)$, po odstránení prvej inverznej transformácie $T^{-1}_{y,\nu}$ na základe definície 2 dostávame

$$P_{p,q}(x, y) = p^{\nu}_0(y) T_{y,\nu} \left(T^{-1}_{x,\mu} T_{x,\mu} P_{p,q}(x, y) \right) + \sum_{l=1}^{3} p^{\nu}_l(y) \left(T^{-1}_{x,\mu} T_{x,\mu} P_{p,q}(x, y_l) \right).$$

Odstránením inverznej transformácie $T^{-1}_{x,\mu}$ opätovnou aplikáciou definície 2 a následným využitím lemy 2 na odstránenie transformácie $T_{x,\mu}$ sa uvedený polynóm zmení nasledovne

$$P_{p,q}(x, y) = p^{\nu}_0(y) T_{y,\nu} \left[p^{\mu}_0(x) \left(P_{p,q}(x_0, y) + z^{\mu}_1(x) A^{\mu}_{p-4}(x, y) \right) + \sum_{k=1}^{3} p^{\mu}_k(x) P_{p,q}(x_k, y) \right] +$$

$$\sum_{l=1}^{3} p^{\nu}_l(y) \left[p^{\mu}_0(x) \left(P_{p,q}(x_0, y_l) + z^{\mu}_1(x) A^{\mu}_{p-4}(x, y_l) \right) + \sum_{k=1}^{3} p^{\mu}_k(x) P_{p,q}(x_k, y_l) \right].$$

V ďalšej úprave sa využíva linearita - dôležitá vlastnosť transformácie $T_{y,\nu}(c_1 F(x, y) + c_2) = c_1 T_{y,\nu} F(x, y) + c_2$ a $A^{\mu}_{p-4}(x, y)$ z lemy 2

$$P_{p,q}(x, y) = p^{\nu}_0(y) \left[p^{\mu}_0(x) \left(T_{y,\nu} P_{p,q}(x_0, y) + z^{\mu}_1(x) \sum_{i=4}^{p}\sum_{j=0}^{q} a_{ij} S^{\mu}_{i-4,4}(x) T_{y,\nu} y^j \right) + \sum_{k=1}^{3} p^{\mu}_k(x) T_{y,\nu} P_{p,q}(x_k, y) \right] +$$

$$\sum_{l=1}^{3} p^{\nu}_l(y) \left[p^{\mu}_0(x) \left(P_{p,q}(x_0, y_l) + z^{\mu}_1(x) \sum_{i=4}^{p}\sum_{j=0}^{q} a_{ij} S^{\mu}_{i-4,4}(x)\ y_l^j \right) + \sum_{k=1}^{3} p^{\mu}_k(x) P_{p,q}(x_k, y_l) \right].$$

Transformáciu $T_{y,\nu}$ z troch miest predchádzajúceho výrazu odstránime jednak použitím lemy 3 pre polynómy, ako aj lemy 1 pre mocninové funkcie. Po uvedenej úprave teda získavame

$$P_{p,q}(x,y)= p_0^{\nu}(y)\Bigg\{ p_0^{\mu}(x)\Bigg[\Big(P_{p,q}(x_0,y_0)+z_1^{\nu}(y)\sum_{i=0}^{p}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\ x_0^i\Big)+$$

$$z_1^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x).\Big(y_0^j+z_1^{\nu}(y)S^{\nu}_{j-4,4}(y)\Big)\Bigg]+\sum_{k=1}^{3}p_k^{\mu}(x)\Big(P_{p,q}(x_k,y_0)+z_1^{\nu}(y)\sum_{i=0}^{p}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\ x_k^i\Big)\Bigg\}+$$

$$\sum_{l=1}^{3}p_l^{\nu}(y)\Bigg[\ p_0^{\mu}(x)\Big(P_{p,q}(x_0,y_l)+z_1^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x)\ y_l^j\Big)+\sum_{k=1}^{3}p_k^{\mu}(x)\,P_{p,q}(x_k,y_l)\Bigg].$$

Po jednoduchých aritmetických úpravách dospejeme k nasledujúcej rovnosti

$$P_{p,q}(x,y)=p_0^{\nu}(y)p_0^{\mu}(x)P_{p,q}(x_0,y_0)+p_0^{\mu}(x)p_0^{\nu}(y)z_1^{\nu}(y)\sum_{i=0}^{p}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\,x_0^i+$$

$$p_0^{\nu}(y)p_0^{\mu}(x)z_1^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x)y_0^j+p_0^{\nu}(y)z_1^{\nu}(y)p_0^{\mu}(x)z_1^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x)\,S^{\nu}_{j-4,4}(y)+$$

$$\sum_{k=1}^{3}p_k^{\mu}(x)p_0^{\nu}(y)P_{p,q}(x_k,y_0)+\sum_{k=1}^{3}p_k^{\mu}(x)p_0^{\nu}(y)z_1^{\nu}(y)\sum_{i=0}^{p}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\,x_k^i+$$

$$\sum_{l=1}^{3}p_l^{\nu}(y)p_0^{\mu}(x)P_{p,q}(x_0,y_l)+\sum_{l=1}^{3}p_l^{\nu}(y)p_0^{\mu}(x)z_1^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x)\,y_l^j+$$

$$\sum_{k=1}^{3}\sum_{l=1}^{3}p_l^{\nu}(y)p_k^{\mu}(x)P_{p,q}(x_k,y_l).$$

Vhodným zoskupením koeficientov polynómu dostávame

$$P_{p,q}(x,y)=\sum_{k=0}^{3}\sum_{l=0}^{3}p_k^{\mu}(x)p_l^{\nu}(y)P_{p,q}(x_k,y_l)+z_4^{\nu}(y)\sum_{i=0}^{p}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\sum_{k=0}^{3}p_k^{\mu}(x).x_k^i+$$

$$z_4^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{q}a_{ij}S^{\mu}_{i-4,4}(x)\sum_{l=0}^{3}p_l^{\nu}(y).y_l^j+z_4^{\mu}(x)z_4^{\nu}(y)\sum_{i=4}^{p}\sum_{j=4}^{q}a_{ij}S^{\mu}_{i-4,4}(x)S^{\nu}_{j-4,4}(y)\ .$$

A na záver rozbitím príslušných súm obmedzíme duplicitný výskyt koeficientov

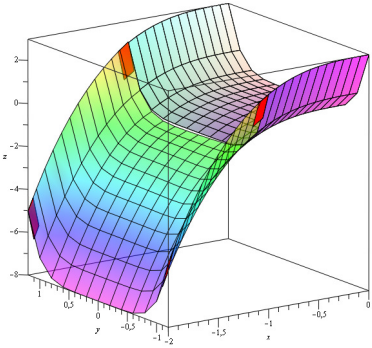
$$P_{p,q}(x,y)=I^{\mu,\nu}_{3,3}(x,y)+z_4^{\nu}(y)\sum_{i=0}^{3}\sum_{j=4}^{q}a_{ij}S^{\nu}_{j-4,4}(y)\sum_{k=0}^{3}p_k(x).x_k^i+z_4^{\mu}(x)\sum_{i=4}^{p}\sum_{j=0}^{3}a_{ij}S^{\mu}_{i-4,4}(x)\sum_{l=0}^{3}p_l(y).y_l^j+$$

$$\sum_{i=4}^{p}\sum_{j=4}^{q}a_{ij}\Bigg(z_4^{\mu}(x)S^{\mu}_{i-4,4}(x).\sum_{l=0}^{3}p_l(y).y_l^j+z_4^{\nu}(y)S^{\nu}_{j-4,4}(y).\sum_{k=0}^{3}p_k(x).x_k^i+z_4^{\mu}(x)z_4^{\nu}(y)S^{\mu}_{i-4,4}(x)S^{\nu}_{j-4,4}(y)\Bigg).$$

□

Vďaka reprezentácii vety 3 sa koeficienty a_{ij}, pre $i=0...p$ a $j=0...q$ rozdelili na štyri skupiny. Pravé horné a ľavé dolné koeficienty sú obsiahnuté v $A^{\nu}_{q-4}(x_k,y)$ a $A^{\mu}_{p-4}(x,y_l)$, kým pravé dolné v $A^{\mu,\nu}_{p-4,q-4}(x,y)$. V reprezentácii sa ľavé horné koeficienty nevyskytujú. V dôsledku reparametrizácie máme namiesto týchto koeficientov šestnásť funkčných hodnôt obsiahnutých v *I*.

Obr. 1 zobrazuje výsledok spojenia dvoch segmentov pomocou vety 3. Horný segment zodpovedá polynómu P(x, y) = x^3+y^6 a dolný je jeho IZA verziou. 16 referenčných bodov sme rozdelili do štyroch skupín po štyri body ležiace blízko k sebe (napr. μ_0=-1, $\mu_1=\mu_0+\tau$, τ = 0.1). Kým horné dve štvorice zohrávajú úlohu pri hladkom spojení daných dvoch segmentov, dolné dve nie – tie by sa mohli použiť pri spojení s tretím segmentom. Horný bod vrchnej štvorice sme úmyselne modifikovali (namiesto $P(\mu_0,\nu_0)$ sme si zvolili $1.02 \cdot P(\mu_0,\nu_0)$, kde μ_0=-1.0 a ν_0=0.4), v dôsledku čoho sa spojitosť medzi dvomi segmentmi narušila – pozri medzeru medzi segmentami.



Obr. 1 IZA spojenie dvoch segmentov pomocou referenčných bodov

4. Záver (Diskusia)

Cieľom tohto článku bolo popísať postup konštrukcie IZA reprezentácie polynómov, odhaliť možnosti jej vylepšenia, čo sa v konečnom dôsledku aj podarilo. Máme pripravenú reprezentáciu, ktorá podľa našich očakávaní nájde svoje využitie pri spájaní segmentov aproximovanej plochy zašumených 3D dát s prihliadnutím na hladký prechod medzi nimi.

5. Literatúra

[1] MATEJČÍKOVÁ A., TÖRÖK Cs., Two-point Transformation and Polynomials, Košice, 2007

[2] TÖRÖK Cs., Reference Points Based Transformation and Approximation, UPJŠ Košice, 2008 (to be published)

[3] SZABÓ I., Discrete transformation of 3D data, Forum Statisticum Slovacum 7/2009, Slovac statistical and demographic society, Bratislava, 2009, p. 169 - 175.

[4] SZABÓ I., TÖRÖK Cs., Reference points based representation of polynomials in space, Management 2010 – Knowledge and management in times of crisis and ensuing development, part II, University of Prešov, 2010, p. 854 – 862.

Kontakt
Imrich Szabó, Csaba Török
P.J. Šafárik University in Košice
Faculty of Science, Institute of Computer science
Jesenná 5, 040 01 Košice, Slovak Republic
imrich.szabo@gmail.com, csaba.torok@upjs.sk

Poďakovanie: *Táto publikácia vznikla vďaka podpore grantového projektu VEGA 1/0131/09 a v rámci OP Výskum a vývoj pre projekt: CaKS - Centrum excelentnosti informatických vied a znalostných systémov (ITMS: 26220120007), spolufinancovaný zo zdrojov Európskeho fondu regionálneho rozvoja..*

Prehľad aktuálne voľne dostupných štatistických softvérov využívaných v štatistickej analýze dát

Overview of current free statistical software used in the data statistical analysis

Mária Szabóová

Abstract: The contribution describes the statistical software, as an important part of statistical analysis, and provides many resources for statistical data analysis. Current overview of free statistical software used in data statistical analysis is given in the first part. The second part includes more details about statistical software using in statistical analysis. Finally, the paper presents advantages and disadvantages of selected statistical softwares.

Key words: statistical software, data statistical analysis

Kľúčové slová: štatistický softvér, štatistická analýza dát

JEL classification: C12

1. Úvod

Štatistický softvér predstavuje komplexný systém, ktorý obsahuje prostriedky pre správu dát a ich analýzu, vizualizáciu a vývoj užívateľských aplikácií. Predstavuje veľmi efektívny prostriedok pri vykonávaní štatistickej analýzy predovšetkým z pohľadu zjednodušenia a urýchlenia výpočtov a grafického zobrazenia. Využíva k tomu rôzne analytické metódy, grafické výstupy, nástroje pre automatizáciu a správu dát. Poskytuje nielen možnosť záznamu dát, ale aj kontrolu dát a ich čistenie. V súčasnosti sú štatistické softvéry veľmi rozšírenou súčasťou využívanou v rámci štatistickej analýzy, a preto sú ponúkané viacerými spoločnosťami. Okrem komerčných štatistických softvérov sa v dnešnej dobe čoraz častejšie vyskytuje aj možnosť využitia voľných, ako aj čiastočne voľných štatistických softvérov.

2. Voľne dostupné štatistické softvéry

Štatistické softvéry poskytujú široký výber základných i pokročilých techník špeciálne vyvinutých a prispôsobených potrebám a požiadavkám jednotlivých používateľov. Uľahčujú vykonávanie rôznych štatistických analýz, ako aj grafických techník analýzy dát. Z hľadiska podpory operačných systémov sú voľne dostupné štatistické softvéri navrhnuté tak, aby ich bolo možné použiť v rámci najbežnejších súčasných operačných systémov.

3. Charakteristika jednotlivých voľne dostupných štatistických softvérov

ADaMSoft predstavuje voľne dostupný štatistický softvér. Umožňuje zhromažďovať a extrahovať dáta z a do rôznych zdrojov. Syntetizuje informácie pomocou štatistických a iných matematických postupov. Softvér má špeciálny dátový archivačný systém, ktorý umožňuje efektívne uchovávať veľké dáta.

Dap poskytuje základné metódy pre správu dát, analýzy a grafiky, ktoré sú bežne používané v štatistickej praxi (základné štatistiky, korelácia a regresia, ANOVA, logistická regresia a iné). Dap je používaný na dátové súbory, ktoré majú veľmi veľa premenných. Ide o štatistický softvér s veľkou flexibilitou programovania.

DATAPLOT je softvérový systém určený pre vedecké vizualizácie a štatistické analýzy. Cieľovou skupinou Dataplot sú užívatelia zaoberajúci sa opisom, modelovaním, vizualizáciou, analýzou, monitorovaním a optimalizáciou vedeckých a technických procesov.

Fityk je program, ktorý ponúka intuitívne grafické rozhranie a tiež rozhranie príkazového riadku, ako aj rôzne metódy optimalizácie.

Tabuľka 1: Prehľad voľne dostupných štatistických softvérov z hľadiska podpory operačných systémov

Štatistické softvéry	Windows	Mac os	Linux	BSD	Unix
ADaMsoft	x	x	x	x	x
Dap	x		x		x
DATAPLOT	x	x	x	x	x
Fityk	x	x	x	x	
GRETL	x	x	x		
JMulti	x	x	x		
MacAnova	x	x	x	x	x
Octave	x		x		x
OpenBugs	x	x	x		x
PINT	x				
R	x	x	x	x	x
SciGraphica		x	x	x	
Scilab	x		x		x
STATPerl	x				
TANAGRA	x				
ViSta	x	x			x

Zdroj: vlastné vypracovanie

Softvér **GRETL** (Gnu Regression, Econometrics and Time-series Library) je voľne dostupný štatistický softvér využívaný v rámci ekonometrickej analýzy.

JMulti je interaktívny softvér určený pre jednorozmerné a viacrozmerné analýzy časových radov. Má grafické používateľské rozhranie, ktoré využíva v rámci štatistických výpočtov.

MacAnova je interaktívny štatistický softvér zahrňujúci mnoho funkcií, zameraný predovšetkým na analýzu rozptylu a súvisiacich modelov, maticovú algebru, analýza časových radov a prieskumné štatistiky. MacAnova môže importovať a exportovať dáta do iných programov.

Octave je primárne určený pre numerické výpočty. Obsahuje rozsiahle nástroje pre riešenie spoločných problémov numerickej lineárnej algebry, hľadanie koreňov nelineárnych rovníc, integrovanie bežných funkcií, diferenciálnych a diferenciálne-algebraických rovníc.

OpenBUGS je široko používaný softvér vhodný pre Bayesovskú analýzu zložitých štatistických modelov pomocou MCMC (Markov Chain Monte Carlo) metódy. Softvér bol navrhnutý tak, aby bol spoľahlivý a účinný v širokom rozsahu testovacích aplikácií.

Pint (Power Analysis In Two-level) je program zameraný na viacúrovňovú analýzu. Je vhodný pre stanovenie štandardnej odchýlky a optimálnej veľkosti vzoriek vo viacúrovňovom modely.

R je štatistický softvér poskytujúci širokú paletu štatistických (lineárne a nelineárne modelovanie, klasické štatistické testy, analýzu časových radov, klasifikácia, zhlukovanie, ...) a grafických techník. Tento softvér je vhodný pre vykonávanie najrôznejších štatistických a grafických techník analýzy dát. Nie je to len štatistický softvér ako taký, ale ide de facto o matematický procesor. Umožňuje testovanie hypotéz, výpočet analýzy rozptylu (ANOVA),

Weibullovo, Studentovo a ďalšie rozdelenia. Obsahuje množstvo základných funkcií a tieto je možné rozšíriť pomocou voľne dostupných balíčkov.

SciGraphica predstavuje vedeckú aplikáciu pre analýzu dát a technickú grafiku. Poskytuje možnosť vykresľovania funkcie pre 2D, 3D a polárne grafy.

Scilab predstavuje softvér využívaný pre numerické výpočty. Hlavnou charakteristickou črtou Scilab je schopnosť zvládnuť matice a zameriava sa tiež na manipuláciu s objektmi, ktoré sú zložitejšie ako numerické matice. Obsahuje algoritmy na riešenie úloh lineárnej algebry a algebry polynómov s možnosťou pracovať nielen v číselnej, ale aj v symbolickej forme.

STATPerl je softvér určený pre rôzne štatistické analýzy, rozdelenia pravdepodobnosti, parametrické a neparametrické testy, analýzu časových radov, korelačné a regresné analýzy, ANOVA, numerická analýza a iné.

TANAGRA predstavuje voľne dostupný softvér určený pre akademické a výskumné účely. Ponúka možnosť využitia rôznych algoritmov, najmä interaktívne a vizuálne konštrukcie rozhodovacích stromov. Tangra taktiež obsahuje faktorové analýzy, parametrické a neparametrické štatistiky, pravidlá asociácie, funkcie pre výber a vytvorenie algoritmov.

ViSta predstavuje špeciálny štatistický softvér využívaný v oblasti vizualizácie výsledkov štatistických analýz.

4. Prehľad výhod a nevýhod vybraných voľne dostupných štatistických softvérov

DATAPLOT

Výhody:

- umožňuje vykonávať matematické a štatistické analýzy,
- určený pre lineárne a nelineárne modelovanie, analýzu rozptylu, t-testy,
- má matematické schopnosti, napr. generovanie náhodných čísel, trigonometrické funkcie, distribučné funkcie, funkcia hustoty pravdepodobnosti a rôzne ďalšie,
- v prípade vizualizácie výsledkov podporuje množstvo formátov a možností, a je tak flexibilný a výkonný,
- poskytuje veľké množstvo špecializovaných grafických formátov a umožňuje generovanie 2D grafov,
- podpora v rámci najbežnejších operačných systémov,
- manuál pre používateľa na stránkach štatistického softvéru DATAPLOT.

Nevýhody:

- neumožňuje tieňovanú 3D grafiku.

GRETL

Výhody:

- voľne šíriteľný štatistický softvér využívaný v rámci ekonometrickej analýzy,
- flexibilný a ľahko použiteľný pri ekonometrických výpočtoch a do určitej miery aj vo vedeckej oblasti,
- pohodlné užívateľské rozhranie,
- jednoduchá obsluha pomocou príkazového riadku,
- vysoká miera údajovej kompatibility softvéru Gretl s inými softvérmi,
- manuál pre používateľa na stránkach štatistického softvéru formulovaný tak, aby prezentoval používané ekonometrické techniky a postupy na konkrétnych príkladoch.

Nevýhody:

- podpora len operačných systémov Linux a Windows, neumožňuje podporu operačných systémov Mac OS, BSD, Unix,
- neobsahuje všetky odhady a testy, ktoré by mohli byť vyžadované v rámci ekonometrie,
- možnosti exportu dát sú v porovnaní s ostatnými softvérmi obmedzenejšie.

MacAnova

Výhody:

- interaktívny, programovateľný počítačový program zameraný na štatistickú analýzu,
- určený pre lineárne modelovanie, viacrozmernú analýzu a analýzu časových radov,
- obsahuje príkazy na vkladanie, upravovanie, kombinovanie a výber dát,
- obsahuje funkcie pre štandardné štatistické výpočty, ako sú popisné štatistiky, P-values, t-testy,
- v rámci premenných môžu byť pripojené krátke popisné poznámky,
- obsahuje príkazy pre vykreslenie dát a výsledkov,
- manuál pre používateľa na stránkach štatistického softvéru MacAnova.

Nevýhody:

- softvér neobsahuje také typy premenných ako ostatné štatistické softvéri,
- má veľmi odlišné funkcie pre čítanie a zápis súborov,
- neobsahuje koncepciu grafických zariadení.

Octave

Výhody:

- poskytuje užívateľovi pohodlné rozhranie pre riešenie lineárnych a nelineárnych problémov a na vykonávanie iných numerických experimentov,
- svojimi funkciami je veľmi podobný komerčnému softvéru Matlab, ktorý sa vyznačuje vysokou cenou,
- navrhnutý tak, aby umožnil robiť náročné výpočty rýchlo a ľahko,
- vyznačuje sa jednoduchou programovateľnosťou,
- manuál pre používateľa na stránkach štatistického softvéru Octave,

Nevýhody:

- softvér nemá vlastný editor skriptov (.m súborov),
- mierne zaostáva v oblasti grafiky, pomocou tohto softvéru nie je možné vytvoriť program s grafickým rozhraním,
- neobsahuje nástroj na tvorbu používateľského rozhrania,
- chýba aj nástroj na modelovanie udalostných systémov.

Štatistický softvér R

Výhody:

- ľahko a rýchlo dostupný softvér,
- podpora rôznych metód,
- umožňuje vytvorenie vlastných funkcií,
- množstvo analytických nástrojov, t. j. okrem základných funkcií obsahuje aj množstvo doplňujúcich balíčkov s funkciami pre rôzne druhy analýz,
- množstvo nielen štandardných, ale aj moderných grafických výstupov,
- na stránkach štatistického softvéru R sú pre používateľa k dispozícii rôzne manuály uľahčujúce prácu s uvedených softvérom,
- identifikácia a odstraňovanie vzniknutých chýb a úprava výstupov,
- pridávanie rôznych obmedzení,
- umožňuje vytvárať najrôznejšie grafy, ktoré sú v mnohých prípadoch dostupné len v komerčných štatistických softvéroch,
- z hľadiska rýchlosti a presnosti počítania je porovnateľný s najlepšími komerčnými štatistickými softvérmi.

Nevýhody:

- používateľ musí vykonať všetky operácie sám, zatiaľ čo v iných aplikáciách je možné danú úlohu vyriešiť len jedným kliknutím.

5. Záver

Štatistické analýzy sa v mnohých prípadoch nezaobídu bez použitia štatistického softvéru. V súčasnosti existuje množstvo komerčných, ale aj voľne dostupných štatistických softvérov. Medzi najznámejší a čoraz viac používaný voľne dostupný štatistický softvér v oblasti vzdelávania, výskumu a vedy patrí softvér R. Jeho výhodou je, že jeho používanie nevyžaduje od bežného používateľa programovacie schopnosti. Program umožňuje vykonávanie rôznych prognóz, analýz a konštrukciu profesionálnych grafov. Ide o systém obsahujúci prostriedky pre správu dát, ich analýzu a vizualizáciu. V súčasnosti patrí v oblasti štatistickej analýzy medzi najrozšírenejší štatistický softvér, a to predovšetkým z dôvodu jeho dostupnosti a poskytovania širokého výberu základných a pokročilých techník pre bežného používateľa potrebných pri výkone najrôznejších štatistických analýz.

Cieľom príspevku zameraného na prehľad aktuálne voľne dostupných štatistických softvérov využívaných v štatistickej analýze dát bolo poukázanie na skutočnosť, že štatistické analýzy je možné vykonávať nielen v komerčných štatistických softvéroch, pri ktorých je problémom finančná nákladovosť a fakt, že bežný užívateľ si tieto softvéry nemôže dovoliť, ale aj možnosť použitia voľne dostupného štatistického softvéru. Výhodou využitia voľne dostupného štatistického softvéru je predovšetkým to, že umožní bežnému užívateľovi robiť štatistické analýzy bez nutnosti vynakladania vysokých výdavkov za programy, ktoré vykonávanie týchto štatistických analýz umožňujú.

Literatúra

[1] Hatrák, M.: Ekonometria. Bratislava: Iura Edition, 2007, 502 s. ISBN 978-80-8078-150-7
[2] Weisberg, S.: Applied Linear Regression. 3rd edition, 2005, 336 p. ISBN 978-04-7166-379-9
[3] http://statistiksoftware.com/free_software.html [cit. 2010-10-08]
[4] http://adamsoft.caspur.it/English/Default.html [cit. 2010-10-08]
[5] http://scigraphica.sourceforge.net/ [cit. 2010-10-09]
[6] http://www.r-project.org/ [cit. 2010-10-09]
[7] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html [cit. 2010-10-09]
[8] http://gretl.sourceforge.net/ [cit. 2010-10-09]
[9] http://www.visualstats.org/ [cit. 2010-10-10]
[10] http://www.stat.umn.edu/macanova/ [cit. 2010-10-10]
[11] http://www.openbugs.info/w/ [cit. 2010-10-11]
[12] http://www.gnu.org/software/octave/ [cit. 2010-10-11]
[13] http://www.scilab.org/ [cit. 2010-10-13]
[14] http://www.statpages.org/miller/openstat/ [cit. 2010-10-13]
[15] http://www.jmulti.com/ [cit. 2010-10-18]
[16] http://www.unipress.waw.pl/fityk/ [cit. 2010-10-18]
[17] http://stat.gamma.rug.nl/multilevel.htm#progPINT [cit. 2010-10-18]
[18] http://www.itl.nist.gov/div898/software/dataplot/homepage.htm [cit. 2010-10-22]

Adresa autora (-ov):

Maria Szabóová, Ing. - externá doktorandka na Katedre financií
Slovenská poľnohospodárska univerzita, Fakulta ekonomiky a manažmentu
Trieda Andreja Hlinku 2, 949 76 Nitra
szaboova.m@gmail.com

Neparametrické metódy odhadu pravdepodobnosti hustoty rozdelenia
Nonparametric estimation method of probability density function

Alena Tartaľová

Abstract: The paper deals with the probability density estimation for household's income in Slovak republic. The classical parametric model is not suitable for the highest values on the right tail. The new approach, nonparametric estimation using kernel density is introduced. We have showed that the free statistical software XploRe for the kernel density estimation is suitable.

Key words: density estimation, kernel density, distribution models, the income of households

Kľúčové slová: odhad hustoty, jadrová hustota, pravdepodobnostné modely, príjem domácností

JEL classification: C13**,** C16 , O15

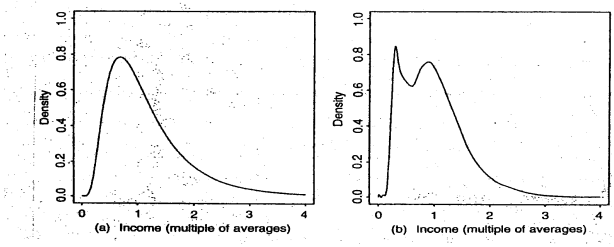
1. Úvod

Znalosť pravdepodobnostného rozdelenia náhodnej premennej z rozličných oblastí predstavuje východisko matematicko-štatistických analýz. Z funkcie hustoty rozdelenia možno nie len vypočítať základné charakteristiky ako sú napríklad stredná hodnota a rozptyl náhodnej premennej, ale možno nájsť aj pravdepodobnosť, že náhodná premenná nadobudne hodnotu z určitého intervalu.

2. Modelovanie príjmov pravdepodobnostnými rozdeleniami

Z teórie pravdepodobnosti poznáme dva základné prístupy k odhadu hustoty výberových údajov: Parametrický prístup, ktorý je v súčasnosti najviac používaný a menej známy neparametrický prístup.

Pri parametrickom prístupe na základe vlastností výberového súboru urobíme predpoklad, že údaje pochádzajú z nejakého známeho rozdelenia s funkciou hustoty $f(x;\theta)$. Následne odhadujeme parametre rozdelenia pomocou jednej zo známych metód (napr. metódou maximálnej vierohodnosti). Problémom pri parametrickom modelovaní je, že obmedzenie odhadu na konkrétny typ pravdepodobnostného rozdelenia urobí určité dôležité predpoklady o hodnotách, ktoré sme nepozorovali, ale ktorých pravdepodobnosť výskytu chceme vypočítať. Na obrázku 1, môžeme vidieť, ako sa líšia odhady hustoty rozdelenia v prípade lognormálneho (parametrického) rozdelenia a neparametrického odhadu hustoty.



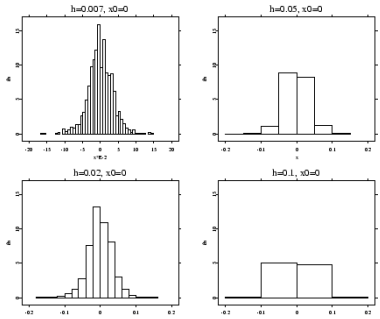
Obrázok 4: Názov obrázku alebo grafu (Times new roman 12, bold, Italic, center)

Hlavný princíp neparametrického prístupu je zhrnutý v citáte: ***„Let the data speak for themselves."*** (pozri [12]), ktorý možno voľne preložiť ako „Nechajme dáta hovoriť samých za seba". To znamená, že nemáme žiaden predpoklad o rozdelení, z ktorého údaje pochádzajú, pracujme iba z nameranými údajmi. Aj pri tejto metóde môžeme naraziť na problémy, ako sú napr. vlastnosti takýchto odhadov. Neparametrických metód odhadu hustoty rozdelenia poznáme z literatúry niekoľko (pozri napr. [4], [5], [9], [12]), sú to napr. histogram, naivný odhad, jadrová hustota, metóda najbližšieho suseda, adaptívny jadrový odhad, metódy ortogonálnych radov a pod.

3. Odhad hustoty rozdelenia pomocou jadrovej hustoty

Najjednoduchšou neparametrickou metódou odhadu hustoty je histogram. Histogram poskytuje prvotnú aproximáciu rozdelenia a na základe jeho tvaru môžeme urobiť predpoklady o hľadanom pravdepodobnostnom rozdelení.

Histogram skonštruujeme tak, že výberové údaje zatriedime do podintervalov a určíme početnosť intervalov, t.j. spočítame koľko výberových údajov padne do nejakého intervalu. Ako dobre vieme, a je to vidieť aj na obrázku 2, tvar histogramu sa mení v závislosti od počtu tried, dĺžky podintervalov a začiatočného bodu. To znamená, že pre nejaké delenie sa môže rozdelenie pravdepodobnosti výberových údajov zdať ako symetrické, pre iné delenie ako nesymetrické, prípadne ako unimodálne alebo duomodálne rozdelenie. Navyše histogram nie je spojitou funkciu, pretože na hraniciach obsahuje skoky. Práve tieto nevýhody konštrukcie histogramu boli podnetom pre rozvoj neparametrických metód odhadu hustoty pomocou tzv. jadrovej funkcie (preklad z angl. kernel density)



Obrázok 2: Histogram rozdelenia pre rôznu šírku a počet intervalov

Odhad hustoty rozdelenia pomocou jadrovej funkcie je založený na podobnej myšlienke ako histogram. Vhodným spôsobom ako odhadnúť funkciu hustoty *f(x)* v bode *x* je vypočítať:

$$f(x) \approx \frac{1}{n.h} \#\{\text{pozorovania, ktoré padnú do intervali okolo hodnoty x}\} \qquad (1)$$

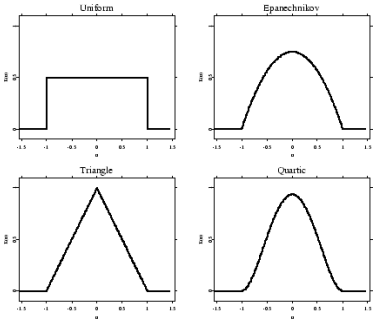
Tento princíp sa dá zovšeobecniť, tak, že pozorovaniam, ktoré sú bližšie k hodnote *x* priradíme väčšiu váhu a pozorovaniam, ktoré sú vzdialenejšie menšiu váhu. Môžeme tak urobiť použitím vhodnej jadrovej funkcie *K*.

Všeobecné vyjadrenie odhadu hustoty rozdelenia $\hat{f}_h(x)$ pre hodnotu funkcie *f(x)* v bode *x*, založenom na náhodnom výbere $x_1, x_2, \ldots, x_n$ pomocou jadrovej funkcie

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (2)$$

Kernel	$K(u)$
Uniform	$\frac{1}{2}\boldsymbol{I}(\lvert u\rvert \le 1)$
Triangle	$(1-\lvert u\rvert)\boldsymbol{I}(\lvert u\rvert \le 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)\boldsymbol{I}(\lvert u\rvert \le 1)$
Quartic	$\frac{15}{16}(1-u^2)^2\boldsymbol{I}(\lvert u\rvert \le 1)$
Triweight	$\frac{35}{32}(1-u^2)^3\boldsymbol{I}(\lvert u\rvert \le 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$
Cosinus	$\frac{\pi}{4}\cos(\frac{\pi}{2}u)\boldsymbol{I}(\lvert u\rvert \le 1)$

Tabuľka 1: Jadrové funkcie



Obrázok 3: Grafy jadrových funkcií

Odhad jadrovej hustoty rozdelenia teda závisí od dvoch parametrov. Pre výpočet jadrovej hustoty je dôležité na začiatku zvoliť jadrovú funkciu *K* a vyhladzujúci parameter *h*. Ukazuje sa, že voľba jadrovej funkcie nie je rozhodujúca (vo viacerých vedeckých článkoch používajú Epanechnikovu funkciu), dôležitejšie je správne zvoliť parameter *h*. Veľmi vysoké hodnoty *h* spôsobujú tzv. prehladenie (angl. oversmoothed) a naopak veľmi nízke hodnoty, tzv. podhladenie (angl. undersmoothed). Pre odhad vyhladzovacieho parametra podľa [12] poznáme dve metódy:

- Metódy typu „plug-in“, ktorých použitie vyžaduje predpoklad o niektorých vlastnostiach neznámeho rozdelenia,
- Odhad založený na metóde krížovej validácie (angl. Cross Validation method)

4. Modelovanie príjmov domácností

Analýza príjmov obyvateľstva je základom rozhodovania v oblasti rozpočtovej a sociálnej politiky. Preto je podkladom ku komplexnej analýze životnej úrovne obyvateľstva. Zisťovaniu životnej úrovne obyvateľstva a sociálnej situácii jednotlivcov je venovaných množstvo publikácií v časopisoch a zborníkoch (domácich aj zahraničných), pozri napr. [1], [8], [10], [11] .

Základom analýzy sú údaje o celkovom disponibilnom príjme domácností pochádzajúce z výberového zisťovania o príjmoch a životných podmienkach domácností (EU SILC 2009).

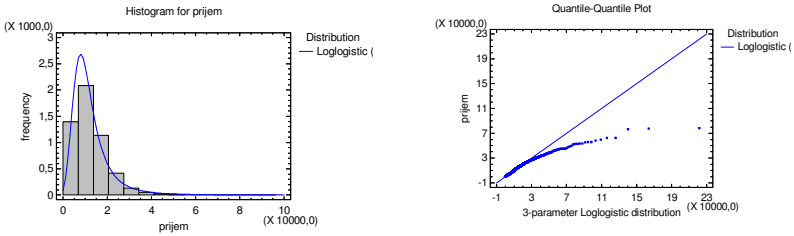
Z výberových charakteristík (v Tabuľke 1) vyplýva, že príjmy domácností sú nesymetrické, pravostranne zošikmené a z histogramu je vidieť, že rozdelenie má predĺžený pravý koniec rozdelenia. Preto ako vhodný model môže poslúžiť niektoré z pravostranne zošikmených rozdelení.

Najprv sme údaje modelovali pomocou známych pravdepodobnostných rozdelení, ktoré sú k dispozícii v programe STATGRAPHICS Centurion XV. Po aplikácii testov dobrej zhody ani jedno z rozdelení, nemodeluje empirické údaje dobre, pre všetky rozdelenia, by sme hypotézu o zhode s teoretickým rozdelením zamietli. Najlepším modelom by z týchto rozdelení mohlo byť trojparametrické loglogistické rozdelenie s odhadnutými parametrami: μ *= 11323,8,* σ^2*= 0,332914*, γ*= -1080,95*, pre ktoré je *p*-hodnota = *2,38624E-8*, čo jemaximálna spomedzi všetkých rozdelení. Na obrázku 3 je však vidieť, že toto rozdelenie nie je vhodným

modelom najmä pre najvyššie príjmy na pravom konci rozdelenia. Vhodný tvar rozdelenia pravdepodobnosti príjmov je teda potrebné nájsť použitím iných metód. Vhodnou alternatívou je použitie kvantilových funkcií, ktorým sa na Slovensku a v Čechách venujú autori napr. v [1], [3], [6]. My v tomto príspevku ukážeme použitie jadrovej hustoty.

Count	5264
Average	12109,5
Standard deviation	7877,6
Coeff. of variation	65,0531%
Minimum	0,0
Maximum	78431,7
Range	78431,7
Stnd. skewness	52,8585
Stnd. kurtosis	92,0941

Tabuľka 2: Výberové charakteristiky



Obrázok 3: Grafické porovnanie zhody s trojparametirickým loglogistickým rozdelením

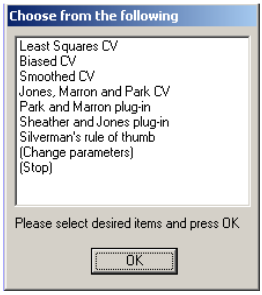
5. Odhad jadrovej hustoty v programe XploRe

Odhad hustoty rozdelenia nájdeme pomocou jadrových odhadov v programe XploRe. Štatistický program XploRe je voľne dostupný na http://fedc.wiwi.hu-berlin.de/xplore.php. Najskôr je potrebné načítať knižnice programu zadaním príkazov:

```
library("smoother")
library("plot")
```

Príkazom **`denbwsel(prijmy)`** sa otvorí ponuka, kde je na výber sedem rôznych metód odhadu parametra *h*. Ako jadrovú funkciu sme zvolili Epanechnikovu jadrovú funkciu. Hodnoty parametra *h* sú v tabuľke 3. Podobne by sme dostali odhady parametra aj pre ďalšie jadrové funkcie, ak by sme zmenili parametre pomocou „(Change parameters)". Záver urobíme na základe grafického porovnania.

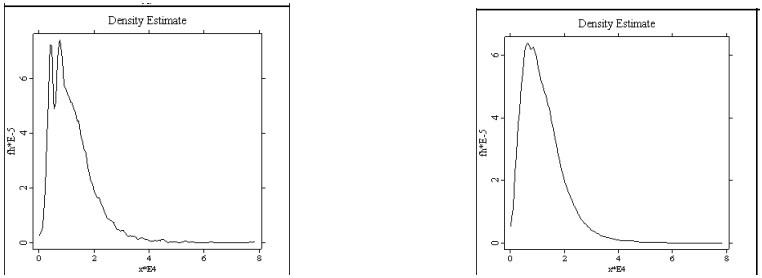
Porovaním odhadu jadrovej hustoty a histogramu na obrázku 5 je zrejmé, že želaný tvar hustoty rozdelenia príjmov najlepšie vystihuje krivka jadrovej hustoty s vypočítaným prametrom *h=3328,57* podľa Silvermanovho pravidla. Metódou krížovej validácie sme dostali krivku, ktorá nie je dostatočne hladká.



Obrázok 4: Metódy odhadu parametra h

Least Squares CV	1176,48
Biased CV	2628,59
Smoothed CV	1369,5
Jones, Marron,Park CV	1369,5
Park, Marron plug-in	1225,14
Silver□an's rule of thumb	3328,57

Tabuľka 3: Hodnoty parametra h



Obrázok 5: Odhad jadrovej hustoty použitím Silvermanovho pravidla (vľavo) a metódy krížovej validácie (vpravo)

x	*Jadrová hustota*	*Trojparametrické loglogistické rozdelenie*
10000	5.7926e-05	6,77e-05
20000	2.03e-05	1,65e-05
30000	4.2861e-06	4,24e-06
50000	1.9103e-07	6,23e-07

Tabuľka 4: Porovnanie neparametrického a parametrického modelu

Pre výšku disponibilného príjmu sme ako vhodný parametrický model zvolili trojparametrické loglogistické rozdelenie. V tabuľke 4 sme vypočítali hodnoty hustoty rozdelenia pre parametrický a neparametrický odhad. Hodnoty jadrovej funkcie dostaneme pomocou príkazu **`fh=denxest(prijmy, 3328.57,"epa", x),`** kde *x = 10000, 20000, 30000* a *50000*. Hodnoty hustoty rozdelenia trojparametrického loglogistického rozdelenia sme vypočítali v programe STATGRAPHICS.

6. Záver

V príspevku sme sa venovali pravdepodobnostnému modelovaniu hustoty rozdelenia náhodnej premennej. Na príklade modelovanie príjmov domácností sme ukázali, že použitie

parametrických modelov je pre heterogénne populácie nepostačujúce a je nutné použiť iné metódy. Použitie neparametrických metód je výhodné vtedy, ak dopredu nepoznáme tvar a vlastnosti rozdelenia, pretože nevyžaduje žiaden predpokladom. Problémom teda ostáva praktická aplikácia týchto metód, pretože väčšina dostupných štatistických balíkov odhad jadrovej hustoty neobsahuje. V programe STATGRAPHICS je jadrová hustota v ponuke *Analyze – Density trace*, ale bez možnosti zmeny parametrov a na výber sú iba dve jadrové funkcie *Boxcar* a *Cosine*. V príspevku sme predstavili štatistický program XploRe, ktorého najväčšou výhodu je, že je voľne dostupný. Obsahuje širokú ponuku rôznych štatistických metód, spolu s názorným pomocníkom, takže sa dá veľmi jednoducho použiť.

7. Literatúra

[1] BARTOŠOVÁ, J. 2009. Výberové šetrění příjmu domácností v České republice, In: Forum Statisticum Slovacum, č.7, 2009, s. 4-9

[2] HARDLE,W. et al.. 2004Nonparametric and Semiparametric Models,[online]. Springer Verlag 2004,[cit. 2010.11.01].
Dostupné na internete : http:// fedc.wiwi.hu- berlin.de/xplore/ebooks/html/spm

[3] PACÁKOVÁ, V., SIPKOVÁ, Ľ., SODOMOVÁ, E. 2005. Štatistické modelovanie príjmov domácností v Slovenskej republike, In: Ekonomický časopis. 2005, č. 4, d. 427- 439.

[4] SIMONOFF, J.S. 1998: Smoothing methods in statistics, Springer-Verlag New York, 1998, 356s. ISBN 0-378-94716-7

[5] SILVERMAN, B. W. 1996 : Density estimation for Statistics and Data Analysis, New York: CHAPMAN AND HALL, 1996, 76 S. ISBN 978-0412246203

[6] SIPKOVÁ, Ľ. , SODOMOVÁ, E. 2007: Modelovanie kvantilovými funkciami, Vydavateľstvo EKONÓM, 2007, 175 s. ISBN 978-80-225-2346-2

[7] SKŘIVÁNKOVÁ, V. , TARTAĽOVÁ, A. 2008: Catastrophic risk management in non-life insurance, In: Ekonomie a Management E+M. ISSN 1212-3609, vol.11, no.2,2008, s. 65-72

[8] STANKOVIČOVÁ, I. 2009. Analýza monetárnej chudoby v domácnostiach Českej republiky, In: Forum Statisticum Slovacum, č.7, 2009, s. 151-156

[9] TSYBAKOV, A. B. 2009: Introduction to Nonparametric Estimation, Springer, 2009,214 s. ISBN 978-0-387-79051-0

[10] ŽELINSKÝ, T. 2010. Porovnanie alternatívnych prístupov k odhadu individuálneho blahobytu domácností ohrozených rizikom chudoby. In: Ekonomický časopis. - ISSN 0013-3035. - Roč. 58, č. 3 (2010), s. 251-270.

[11] ŽELINSKÝ, T. 2010. Analýza chudoby na Slovensku založená na koncepte relatívnej deprivácie. In: Politická ekonomie. - ISSN 0032-3233. - Vol. 58, no. 4 (2010), p. 542-565.

[12] WAND, M. P., JONES, M.C. 1995: Kernel Smoothing, London: Chapman and Hall, 1995, 224 s. ISBN 978-0412552700

Adresa autora:

Alena Tartaľová, Mgr.
Technická univerzita v Košiciach, Ekonomická fakulta
Katedra aplikovanej matematiky a hospodárskej informatiky
Nemcovej 32, 040 01 Košice
alena.tartalova@tuke.sk

Príspevok bol vytvorený s podporou vedeckovýskumného projektu VEGA 1/0370/08 „Regionálny prístup k meraniu sociálneho kapitálu a chudoby“

Analýza elektrotechnického priemyslu v roku 2009 pomocou pomerových ukazovateľov ekonomiky firmy v absolútnom vyjadrení

The analysis in the electronics industry in 2009 via proportional indicators of economy firm in absolute terms

Ondrej Tretiak

Abstract: This paper analyzes the state of the electronics industry in 2009. The aim is to assess how they influence each other individual selected business indicators in electronics industry on the overall development of the electronics industry in this year and stressed its importance for Slovakia. Mention here, what are assets, which reached revenues in 2009, as assets impact to the amount of revenues of companies and other indicators. Individual calculations were made on the basis of statistical data.

Key words: ***Assets, Revenue, Added value, Takings, Excel, Proportion, indicator***

Kľúčové slová: majetok, výnosy, pridaná hodnota, tržby, Excel, pomer, ukazovateľ

1 Úvod

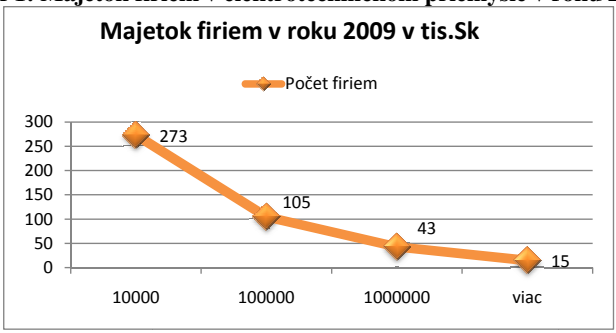
Priemysel na Slovensku zažíva od začiatku deväťdesiatych rokov významnú štrukturálnu zmenu. Veľký vplyv na ňu mala za posledné roky aj hospodárska kríza. Najväčšie zmeny sa prejavili tiež v najvýznamnejšie a zároveň najdynamickejšie sa rozvíjajúcom elektrotechnickom odvetví priemyslu na Slovensku. Toto odvetvie tvorí jeden zo základných opôr priemyslu a zároveň celej slovenskej ekonomiky. Prejavuje sa v ňom veľký rast moderných sofistikovaných podnikov a zároveň zanikajú zastarané prevádzky, čím sa prečisťuje dané odvetvie. Najväčšiu zásluhu na tom majú pobočky zahraničných spoločností, ktoré dokázali na troskách krachujúcich fabrík a niekedy aj na zelenej lúke vytvoriť svoje závody, niektoré dokonca koncernové výrobne ústredia pre celú Európu.

V roku 2009 odvetvie síce mierne padlo, ale pridaná hodnota a rentabilita sa vyšvihli na nové maximá. Prinieslo to však daň, v podobe zvýšení nezamestnanosti v tomto odvetví o 15 %. Možno to pokladať za opačný stav na aký bol elektrotechnický priemysel zvyknutý. Dôvodom bol rast, ktorý tu bol za posledných desať rokov. Počas tohto obdobia tržby slovenských podnikov v tomto odvetví narástli 7 násobne a tvorba pridanej hodnoty približne štvornásobne. Zároveň počet zamestnancov narastal rýchlejšie ako pridaná hodnota, tržby či rentabilita. Podľa údajov Štatistického úradu Slovenskej republiky pôsobilo v roku 2008 pôsobilo v tomto odvetví 182 podnikov s viac ako 20 zamestnancami. V roku 2009 ich počet dosiahol 188, čo ukazuje na príchod nových investorov a rast malých podnikov na väčšie. Podľa celkových tržieb v elektrotechnickom priemysle patrí tomuto odvetviu druhá pozícia za výrobou aut, teda za strojárenským odvetvím.

Z hľadiska budúcnosti vývoja elektrotechnického priemyslu je dôležité, že globálne i regionálne pôsobiace koncerny na Slovensku už nehľadajú len lacnú pracovnú silu a výrobu, ale začínajú hľadať aj podmienky na vytvorenie centier svojho podnikania. Spolu s činnosťami akými sú servis, technická podpora a v sporadických prípadoch aj dizajn, alebo dokonca aj vývoj a výskum. Spolu s kvalitným stupňom elektrotechnických fakúlt slovenských technických univerzít a s elektrotechnickou tradíciou sa ponúkajú nové možnosti na ďalší rozvoj elektrotechnického priemyslu. Ich využitie bude dôležité, keď sa už primárne výhody, ktoré mnohých investorov pritiahli začínajú postupne vytrácať a časť z nich uvažuje nad efektívnosťou ďalšieho pôsobenia na Slovensku.

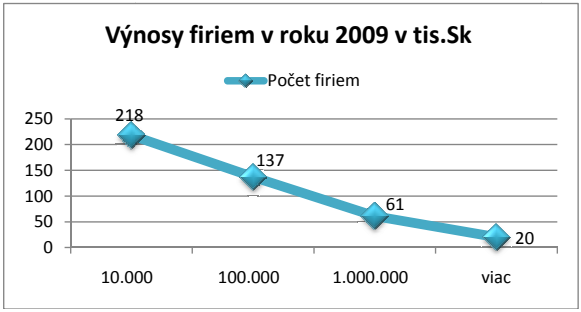
V najväčšej miere dopad krízových opatrení pocítili na konci roku 2008 jednoduchšie výroby elektrotechnického biznisu. Dôvodom je, že sú vo veľkej miere závislé od veľkých odberateľov, ale zároveň sú pod neustálim tlakom racionalizácie. Toto neustále znižovanie nákladov si vyžaduje svoju daň, na ktorú malo vplyv aj zavedenie novej meny na Slovensku. Prijatie eura a jeho silný konverzný kurz malo veľký vplyv na urýchlenie očisťujúceho procesu. Negatívny vplyv to malo hlavne na firmy z vysokým podielom pracovnej sily. Komparatívna výhoda Slovenka – lacná pracovná sila, v pomerne krátkom čase zdražela viac ako sa predpokladalo. Firmy z vysokým podielom ľudskej práce začali postupne opúšťať Slovensko, prípadne znižovať tuzemské počty a kapacity zamestnancov. Takéto odchody zanechali za sebou veľké množstvo kvalifikovaných pracovných síl, ktoré môžu využiť a zamestnať nové elektrotechnické firmy prichádzajúce na Slovensko. [1]

Graf 1: Majetok firiem v elektrotechnickom priemysle v roku 2009



V roku 2009 mali firmy v elektrotechnickom priemysle spolu majetok v hodnote 45 754 869 000 Sk teda 1 518 783 410 eur. Zo 436 firiem 273 disponovalo majetkom v hodnote do 10 000 0000 Sk čo je 333 200 eur, 105 firiem disponovalo majetkom do 100 000 000 Sk , čo predstavovalo 3,32 miliónov eur , 43 firiem malo 100 000 000 Sk, čo bolo 33,2 miliónov eur, a 15 firiem disponovalo majetkom väčším ako 33,2 miliónov eur. [2]

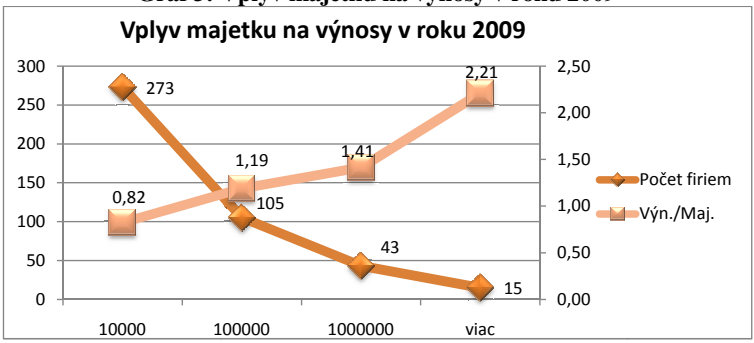
Graf 2: Výnosy firiem v elektrotechnickom priemysle v roku 2009



V roku 2009 zo 436 firiem pôsobiacich v elektrotechnickom priemysle dosiahlo výnosy 80 605 313 000 Sk, teda 2,675 miliardy eur, 218 firiem dosiahlo výnosy do10 000 000 Sk, čo predstavuje sumu 332000 Eur, 137 firiem malo výnosy do 100 000 000 Sk, čo predstavuje 3, 320 milióna eur , 61 firiem dosiahla výnosy do 100 000 000 Sk, čo je suma

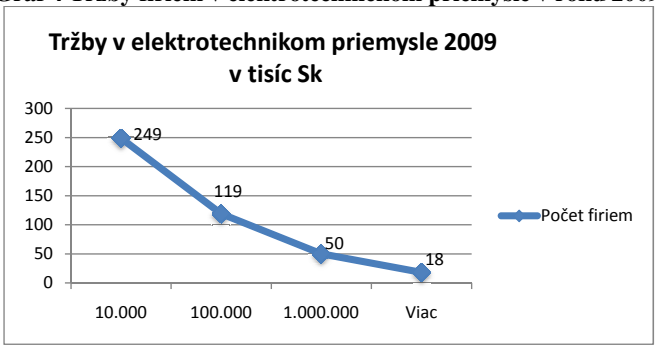
33,2 miliónov eur, a 20 firiem dosiahlo výnosy nad 33,2 miliónov eur. [2]

Graf 3: Vplyv majetku na výnosy v roku 2009



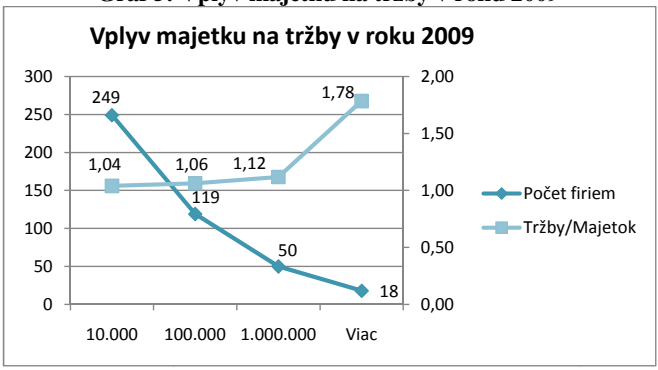
V roku 2009 najväčšie výnosy v elektrotechnickom priemysle dosahovali firmy z najvyšším majetkom, teda zo vrastajúcim majetkom narastali aj výnosy. Pätnásť firiem dosiahlo hodnotu výnosov nad 1000 000 000 Sk, teda viac ako 33,2 milióna eur. [2]

Graf 4 Tržby firiem v elektrotechnickom priemysle v roku 2009



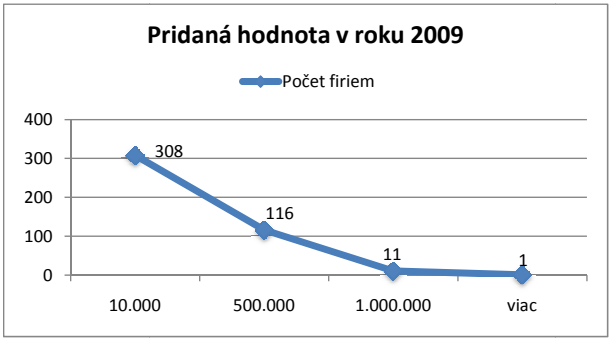
V roku 2009 malo 249 firiem v elektrotechnickom priemysle tržby do 10 000 000 Sk, čo predstavuje 332 000 eur, 119 firiem malo tržby do 100 000 000 Sk, čo predstavuje 3,32 milióna eur, 50 firiem dosahovalo tržbami hodnotu 100 000 000 Sk, teda 33,2 miliónov eur. 18 firiem dosiahlo tržby nad 100 000 000 Sk, čo predstavovalo sumu 33,2 miliónov eur Všetky firmy dosiahli tržby spolu asi 66 miliardy SK, čo predstavuje hodnotu 2,19 miliard eur.[2]

Graf 5: Vplyv majetku na tržby v roku 2009



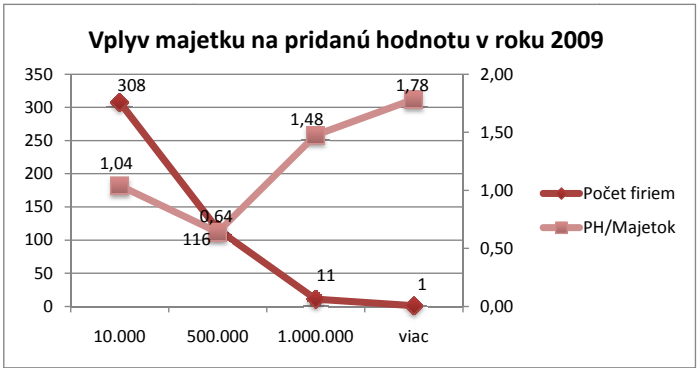
V roku 2009 až 18 firiem dosiahlo tržby nad 33,2 milióna eur. Zo zvyšujúcim sa majetkom firiem narastali aj dosiahnuté tržby, ako je zrejme z grafu. [2]

Graf 6: Pridaná hodnota v roku 2009



V roku 2009 dosiahlo až 308 firiem pridanú hodnotu do 10 000 000 Sk, teda 332 tisíc eur, 116 firiem malo pridanú hodnotu do 100 000 000 Sk 3,32 miliónov eur a 11 firiem dosiahlo pridanú hodnotu do 1000 000 000 Sk, teda 33,2 milióna eur. Len jedna firma presiahla miliardovú pridanú hodnotu, teda sumu nad 33,2 milióna eur. [2]

Graf 7: Vplyv majetku na pridanú hodnotu v roku 2009



V roku 2009 dosahovala väčšina firiem v elektrotechnickom priemysle pridanú hodnotu do až 10 000 000 Sk, teda do 332 tisíc eur. Uvedený graf znázorňuje, že firmy z majetkom do 332 tisíc eur dosahujú oproti firmám z hodnotou majetku 3,32 milióna eur vyššiu pridanú hodnotu. Avšak firmy z majetkom do 33,2 milióna eur a nad túto sumu dosahujú už vyššiu pridanú hodnotu. Z toho vyplýva, že výška majetku podniku nemá rozhodujúci vplyv na ekonomický zisk podniku, teda na ekonomickú pridanú hodnotu. [2]

2 Záver

Z analyzovaných údajov v elektrotechnickom priemysle z roku 2009 vyplýva, že výška majetku podniku významné vplýva na dosahovanie jednotlivých ekonomických ukazovateľov efektívnosti podniku, ktorými boli analyzované hodnoty: výnosy, tržby či pridaná hodnota. Je však zrejmé na základe posledného grafu, že výška majetku nemá rozhodujúci vplyv na dosahovanie zisku podniku, ale určite sa radí k významným a dôležitým faktorom, ktoré zabezpečujú pre podnik rentabilitu a konkurencieschopnosť.

3 Literatúra

[1] http://www.economy.gov.sk/elektrotechnicky-priemysel-5841/127526s (EP_analyza_2009-1.pdf)

[2] http://portal.statistics.sk/showdoc.do?docid=4

Adresa autora :
Ondrej Tretiak, Ing. et Ing.
Gercenova 13
85101 Bratislava
ondrej.tretiak@yahoo.com
ondrejtretiak@gmail.com

Rozvoj štatistického myslenia žiakov prostredníctvom hry
The development of statistic thinking of students through game

Eva Uhrinová

Abstract: Nowadays, we are still more in contact with the information presented through graphic representation. To make such information useful for public, it is important to develop statistic thinking from elementary schools. A game can be a suitable educational method. In the article, we compare the games Understanding Percent and Matching Fractions which can be found freeware on the Internet and can develop competences needed for creating and understanding the pie charts by students. We consider the game Matching Fractions as more interesting for children. However, we recommend including both games into the educational process, specifically for elementary school pupils of 6th grade. We advise to integrate the game Matching Fractions as the first of the games.

Key words: didactic game, pie chart

Kľúčové slová: didaktická hra, koláčový graf

JEL classification: C 18

1. Úvod

Grafické znázorňovanie slúži pre rýchlu a názornú prezentáciu štatistických výsledkov. Aby laická spoločnosť vedela získavať potrebné informácie z týchto štatistických vyjadrení, je potrebné rozvíjať štatistické myslenie už od základnej školy. V súčasnosti sa do popredia dostávajú nové edukačné metódy, ktoré odbúravajú klasický typ vyučovacej hodiny a majú žiakom pomôcť osvojiť si danú problematiku ľahšie, rýchlejšie, zaujímavejšie. Jednou z nich môže byť metóda didaktických hier. Aj Vankúš [1] vo svojej práci zdôrazňuje používanie didaktických hier, ako vhodnej metódy na komplexný rozvoj osobnosti žiaka a aktívnu prácu žiaka na hodinách**.**

V článku sa zameriavame na porovnanie dvoch hier z hľadiska rozvoja štatistických kompetencií u žiakov ZŠ.

2. Materiál a metódy

Hra je podľa viacerých významných odborníkov jednou z **podmienok učenia sa**, pretože dieťa hrou zisťuje, ako veci, javy, ľudia okolo neho fungujú, ale aj zisťuje, čo môže urobiť, a ako veci, javy, ľudí okolo seba môžu vlastnými schopnosťami ovplyvniť. **Hra je efektívny prostriedok učenia**, pretože komplementárne spája všetky nutné podmienky učenia a to: induktívnu skúsenosť, kognitívnu disonanciu (poznávaciu nerovnováhu), sociálnu interakciu, fyzickú skúsenosť a využitie vlastnej kompetentnosti [2]**.** Keďže väčšina detí školského veku trávi svoj čas za počítačom, pokladáme za vhodné predostrieť im problematiku štatistického myslenia nenásilnou formou a to metódou hry – **hry na počítači.**

V súčasnosti môžeme nájsť na internete dostatočné množstvo edukačných hier vhodných na rozvoj matematických kompetencií. Bohatá na matematické hry je napríklad stránka: http://pertoldova.webzdarma.cz/vyuka/matematika/zaci/mat10.htm, kde je možnosť výberu matematických hier podľa problematiky rôznych tematických celkov preberaných v školskej praxi.

Našim cieľom bolo porovnať dostupné hry na internete, ktoré môžu slúžiť na rozvoj štatistickej gramotnosti u žiakov ZŠ. Zvolili sme hry: Understanding Percent a Matching

Fractions, ktoré pokladáme za vhodné na rozvoj kompetencií potrebných pre tvorbu a porozumenie koláčových grafov. Hry sú dostupné na týchto stránkach:

- Understanding Percent: http://www.harcourtschool.com/activity/elab2004/gr5/17.html
- Matching Fractions: http://www.sheppardsoftware.com/mathgames/fractions/ memory_ _fractions3.htm

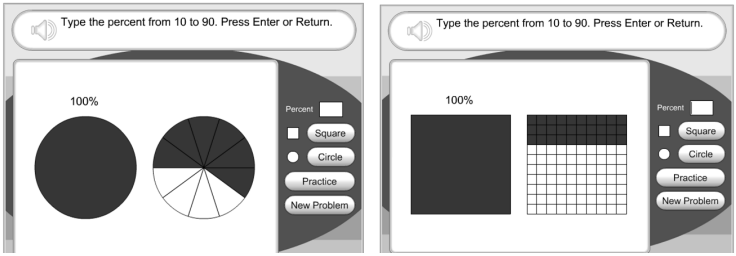
3. Výsledky a diskusia

Hra Understanding Percent

Po spustení hry sa na obrazovke znázornia dva kruhy, alebo dva štvorce (obrázok 1). Jeden znázorňuje 100%, v druhom sa nachádza vyfarbená časť z celku, ktorá znázorňuje istý počet percent. Úlohou hráča je do políčka *Percent* napísať, aká časť v % je na tomto obrázku znázornená. O správnosti a nesprávnosti vyriešenia úlohy nás oboznámi zvukový efekt: - veselý- správne vyriešenie úlohy, - smutný – nesprávne vyriešenie úlohy [5].

Tlačidlo *New Problem* slúži na zobrazenie novej úlohy. Novú úlohu si môžeme zvoliť aj keď sme neodpovedali na predchádzajúcu úlohu. Kliknutím na *New Problem* sa postupne strieda úloha so štvorcom a kruhom a vždy sa zobrazí iná vyfarbená časť z daného útvaru.

Táto hra ponúka aj opačný spôsob postupu- voľbou tlačidla *Practice* a zvolením útvaru štvorca (*Square*), alebo kruhu (*Circle*), môžeme do políčka *Percent* napísať ľubovoľný počet percent, ktorý chceme, aby sa nám na obrázku znázornil.



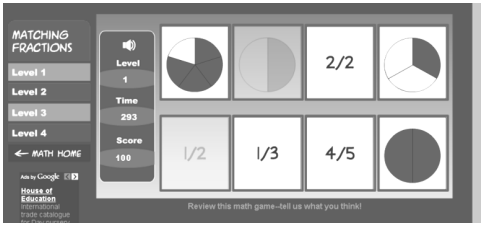
Obrázok 1: Ukážka hry Understanding Percent [5]

Pokladáme za vhodné v edukačnom procese začať funkciou *Practice*, aby žiaci najprv odpozorovali, ako sa zo zadávaním počtu percent, mení vyfarbená časť zvoleného útvaru. Následne, zvolením funkcie *New Problem* si overia, či už k vyfarbenej časti vedia zadať počet percent.

Pre niektorých žiakov môže byť náročné zistiť počet percent zodpovedajúci vyfarbenej časti prostredníctvom kruhu a práve preto, možnosť voľby medzi kruhom a štvorcom, pokladáme za významnú pre lepšie pochopenie a upevnenie danej problematiky.

Hra Matching Fractions

Od predchádzajúcej hry sa líši vyjadrením časti z celku v podobe zlomku a nie percent ako je tomu v predchádzajúcej hre. Táto hra funguje podobne ako pexeso. Úlohou hráča je označiť dvojice kartičiek, ktoré patria k sebe – obrázok a číslo. Keď správne pospájame dané kartičky, posunie nás hra do vyššieho levelu. Hra nám umožňuje zvoliť si ľubovoľný level. Platí - čím vyšší level - tým náročnejšia úloha. Ak označíme správnu dvojicu kartičiek, zaznie veselý tón a kartičky stmavnú (obrázok 2). Ak označíme nesprávnu dvojicu, zaznie smutný tón a kartičky zostanú v pôvodnom stave.



Obrázok 2 Ukážka hry Matching Fractions [6]

Na prejdenie levelu máme istý časový limit, ktorého dĺžka závisí od levelu, v ktorom sa nachádzame. V tejto hre môže hráč získavať aj body – skóre. Skóre získavame za každý správne označený pár kartičiek a aj za zostávajúci čas [6].

Hru Matching Fractions pokladáme za zaujímavejšiu pre deti vzhľadom na to, že žiak v nej získava skóre, ktoré si môže porovnať s ostatnými spolužiakmi. Prípadne si môže porovnávať svoje vlastné skóre a pozorovať tak, ako sa mu darí. Podľa nášho názoru, je vhodné, prejsť so žiakmi obidve hry. Ako prvú by sme zaradili Matching Fractions, v ktorej sa žiaci naučia vyjadriť zlomkom istú časť kruhu a naopak. Po zvládnutí tejto hry žiakmi, navrhujeme pokračovať hrou Understanding Percent, vďaka ktorej sa žiaci naučia vyjadriť percentami vyfarbenú časť kruhu a naopak. Vzhľadom na to, že žiaci už budú vedieť danú vyfarbenú časť kruhu vyjadriť zlomkom a tu sa ju naučiť vyjadriť percentami, intuitívnou formou nájdu vzťah medzi vyjadrením v tvare zlomku a počtom percent.

Zaujímavý by mohol byť aj postup využívania oboch hier naraz

4. Záver

V príspevku ponúkame porovnanie dvoch hier vhodných na rozvoj štatistických kompetencií u žiakov ZŠ. Podľa štátneho vzdelávacieho programu ISCED 2 je súčasťou obsahu tematického okruhu *Kombinatorika, pravdepodobnosť, štatistika* v 6. ročníku grafické znázornenie údajov [7]**.** Vzhľadom na to, je vhodné zaradiť tieto hry do 6. ročníka ZŠ. Zaradenie týchto hier až od 6. ročníka súvisí aj s tým, že so zlomkami sa žiaci oboznamujú až v 6. ročníku. Hoci percentá, ako súčasť tematického okruhu *Čísla, premenná a počtové výkony s číslami*, sú zaradené až do 7. ročníka ZŠ, pokladáme hru Understanding percent vhodnú aj pre 6. ročník, keďže žiaci sa už stretávajú s údajmi vyjadrených v percentách a pre grafické znázornenie údajov do koláčového grafu je vhodné poznať aj percentuálne vyjadrenie jednotlivých údajov.

5. Literatúra

[1]VANKÚŠ, P. 2006. *Efektívnosť vyučovania predmetu matematika metódou didaktických hier*. Autoreferát dizertačnej práce. 2006. Bratislava: Katedre algebry, geometrie a didaktiky matematiky Fakulty matematiky, fyziky a informatiky. [2010-05-03].Dostupné na internete: http://www.ddm.fmph.uniba.sk/files/dizertacky/autoPeto.pdf

[2]KIKUŠOVÁ, S. – KRÁLIKOVÁ, M. 2004. Dieťa a hra. 1. vyd. Bratislava: Sofa, 2004.ISBN 80-89033-42-3

[3]TIRPÁKOVÁ, A. – MARKECHOVÁ, D. 2008**.** *Štatistika v praxi s popisom postupu práce v programe Excel.* 1. vyd. Nitra: FPV UKF, 2008. 390s. ISBN 978-80-8094-283-0

[4]MATEMATIKA NA INTERNETU, 2010 . ZŠ a MŠ Na Beránku Praha. [2010-09-20]. Dostupné na internete: http://pertoldova.webzdarma.cz/vyuka/matematika/zaci/mat10.htm

[5]E-LAB, UNDERSTANDING PERCENT, 2010 . [2010-09-20]. Dostupné na internete:<http://www.harcourtschool.com/activity/elab2004/gr5/17.html>

[6]MATH GAMES: MATCHING FRACTIONS, 2010. [2010-09-20]. Dostupné na internete:<http://www.sheppardsoftware.com/mathgames/fractions/memory_fractions3.htm>

[7]ŠTÁTNY VZDELÁVACÍ PROGRAM, PRÍLOHA ISCED 3A - matematika, Štátny pedagogický ústav BA, 2010.[2010-09-20]. Dostupné na internete: <http://www.statpedu.sk/sk/filemanager/view/723 >

Adresa autora (-ov):

Eva Uhrinová, Mgr.
Katedra matematiky UKF Nitra
Trieda A. Hlinku 1
94974 Nitra
eva.uhrinova@ukf.sk

Porovnanie metód odhadu hraníc subjektívnej chudoby
Comparison of Subjective Poverty Lines Estimation Methods

Tomáš Želinský

Abstract: The aim of paper is to demonstrate estimation of monetary subjective poverty lines using the Slovak EU-SILC 2009 microdata. Two variables are used and three different methods are compared: logistic regression, sensitivity/specificity curves and comparison of cumulative distribution functions. Twelve different monetary subjective poverty lines are obtained and their values differ considerably.

Key words: Poverty, subjective poverty, poverty lines, logistic regression, sensitivity/ specificity curves, EU-SILC.

Kľúčové slová: chudoba, subjektívna chudoba, hranice chudoby, logistická regresia, krivky senzitivity a špecifickosti, EU-SILC.

JEL classification: I32, I33

1. Úvod

Hodnotenie javov subjektívnej povahy býva často problematické, nakoľko výskumník sa musí spoliehať na odpovede respondentov, ktoré môžu byť v procese dopytovania ovplyvnené viacerými faktormi. V praxi nám ale neostáva nič iné, ako spoliehať sa na hodnotenie ľudí, ktorých sa problém bezprostredne týka a vychádzať z predpokladu, že ľudia sú sami najlepšie schopní zhodnotiť kvalitu svojho života [1].

Cieľom príspevku je ilustrovať použitie metód vhodných na odhad (peňažných) hraníc subjektívnej chudoby s využitím údajov EU-SILC 2009 [2] za Slovenskú republiku a následne porovnať výsledky.

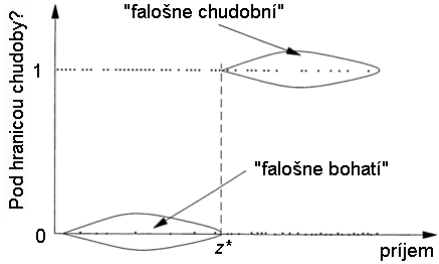
Na odhad hraníc subjektívnej chudoby možno použiť množstvo metód, v článku uskutočníme porovnanie výsledkov získaných s využitím odhadov založených na logistickej regresii, kriviek senzitivity/špecifickosti a distribučných funkciách rozdelení príjmov (pozri napr. [3], [4], [5], [6], [7], [8]). Všetky odhady sú uskutočnené v prostredí programu R.

2. Metódy odhadu hraníc subjektívnej chudoby

V procese odhadu (peňažných) hraníc subjektívnej chudoby na základe údajov EU-SILC je v prvom rade potrebné definovať premennú, na základe ktorej bude hodnotený individuálny status domácnosti z pohľadu vnímania subjektívnej chudoby. Za najvhodnejšie možno považovať premenné *HS130 (minimálne požadovaný čistý mesačný príjem)* a *HS120 (stupeň schopnosti splácať zvyčajné výdavky)*.

Následne je potrebné určiť podmienky, kedy môžeme domácnosť považovať za subjektívne chudobnú. V prípade premennej (*HS130*) možno za chudobné považovať tie domácnosti, ktoré uviedli hodnotu minimálne požadovaného čistého mesačného príjmu vyššiu, ako je mesačný disponibilný príjem domácnosti. V prípade premennej (*HS120*) môžu nastať tri prípady, kedy je domácnosť možné považovať za subjektívne chudobnú: 1. domácnosti schopné splácať zvyčajné výdavky s veľkými ťažkosťami; 2. domácnosti schopné splácať zvyčajné výdavky s ťažkosťami alebo veľkými ťažkosťami; 3. domácnosti schopné splácať zvyčajné výdavky s určitými ťažkosťami, ťažkosťami alebo veľkými ťažkosťami (pre detaily pozri napr. [6]).

V oboch prípadoch je výsledná premenná binárna s kategóriami „*1*“: domácnosť sama seba považuje za subjektívnu chudobnú; „*0*“: domácnosť sama seba nepovažuje za subjektívne chudobnú. Keďže naším cieľom je odhadnúť peňažnú hranicu subjektívnej chudoby, výslednú binárnu premennú týkajúcu sa subjektívnej chudoby je potrebné dať do vzťahu k príjmom domácností (v našom prípade uvažujeme s ekvivalentným disponibilným príjmom domácnosti – premenná *HX100*). Danú situáciu možno schematicky zachytiť (pozri obr. 1).



Obrázok 1: Odhad subjektívnej hranice chudoby

Zdroj: [9, s. 125]

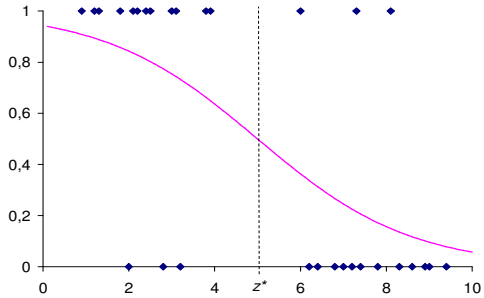
Hodnota hranice subjektívnej chudoby (z^*) je určená ako hodnota, ktorej zodpovedá najvyššia pravdepodobnosť, že odpovede respondentov o ich statuse chudoby korešpondujú so statusom (resp. minimalizujeme počet nesprávne identifikovaných chudobných a nechudobných domácností, t. j. „falošne bohatých“ a „falošne chudobných“), ktorý by sme zistili porovnaním hodnoty z^* s ich príjmom.

V príspevku zameriame pozornosť na odhad subjektívnej peňažnej hranice chudoby podľa vyššie opísaného princípu s použitím logistickej regresie, kriviek senzitivity a špecifickosti a distribučnej funkcie rozdelení príjmov.

2.1 Odhad založený na logistickej regresii

Použitie *logistickej regresie* [10] umožní odhadnúť pravdepodobnosť, že vzhľadom na danú hodnotu ekvivalentného disponibilného príjmu možno domácnosť považovať za subjektívne chudobnú. Majme dvojrozmerné pozorovania $[x_i; y_i]$, kde x_i je ekvivalentný disponibilný príjem i-tej domácnosti a y_i je binárna premenná popisujúca subjektívne vnímanie pocitu chudoby i-tej domácnosti. Máme pravdepodobnosť $P(y_i = 1|x_i)$, že i-tá domácnosť je subjektívne chudobná a $P(y_i = 0|x_i)$, že i-tá domácnosť nie je subjektívne chudobná.

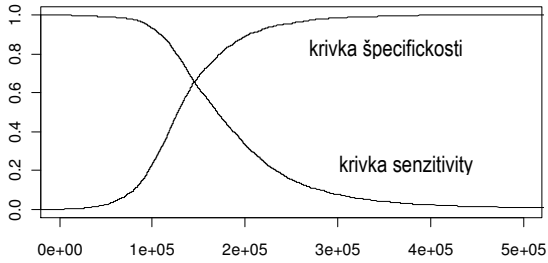
Za subjektívne chudobné budeme považovať tie domácnosti, u ktorých pre odhadnutú pravdepodobnosť platí $\hat{p} \geq 0{,}5$. Hranicu chudoby z^* určíme ako tú hodnotu ekvivalentného disponibilného príjmu x_i, ktorej zodpovedá odhadnutá pravdepodobnosť $\hat{p} = 0{,}5$ (pozri obrázok 2). [5]



Obrázok 2: Využitie logistickej regresie pri odhade subjektívnej chudoby
Zdroj: [5]

2.2 Odhad založený na krivkách senzitivity a špecifickosti

Použitie kriviek senzitivity a špecifickosti vychádza z pravdepodobnostného prístupu k odhadu hraníc subjektívnej chudoby [6]. Krivky senzitivity a špecifickosti sú z tzv. ROC priestoru (ROC – *Receiver Operating Characteristic*). ROC krivky sa pôvodne využívali v teórii detekcie signálu. V prípade, že naším cieľom je určiť optimálny bod zlomu na klasifikačné účely (v našom prípade hodnotu príjmu na klasifikáciu domácností na chudobné/nechudobné), jednou z možností je zvoliť bod, v ktorom dochádza k maximalizácii ako senzitivity, tak aj špecifickosti. Takýmto bodom môže byť priesečník kriviek senzitivity a špecifickosti [11], pozri obr. 3.



Obrázok 3: Priesečník kriviek senzitivity a špecifickosti
Zdroj: vlastný obrázok

S využitím uvedeného prístupu je odhad subjektívnej hranice chudoby z^* daný takou hodnotou príjmu x_k, pre ktorú platí:

$$\mathrm{P}\left(y_j = 1 \mid x_j \le x_k\right) - \mathrm{P}\left(y_j = 0 \mid x_j > x_k\right) = 0\,, \qquad (1)$$

kde

x_j je príjem j-tej domácnosti,

x_k je možný bod zlomu, t. j. potenciálna hranica chudoby, $x_k \in \langle x_{\min}, x_{\max}\rangle$; $k = 1, 2, \ldots, n$; n je veľkosť vzorky,

$\mathrm{P}(y_j = 1 \mid x_j \le x_k)$ je pravdepodobnosť, že náhodne vybratá domácnosť subjektívne vníma svoju situáciu ako stav chudoby, za predpokladu, že jej príjem je nanajvýš rovný príslušnej potenciálnej hranici chudoby x_k,

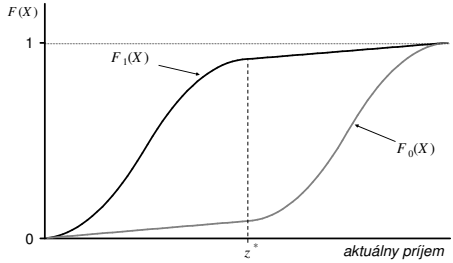
$P(y_j = 0 \mid x_j > x_k)$ je pravdepodobnosť, že náhodne vybratá domácnosť subjektívne nevníma svoju situáciu ako stav chudoby, za predpokladu, že jej príjem je vyšší ako príslušná potenciálna hranica chudoby x_k.

2.3 Odhad založený na distribučných funkciách rozdelení príjmov

Nech rozdelenie príjmov osôb, ktoré uviedli, že sa cítia byť chudobné ($y_i = 1$), je dané distribučnou funkciou $F_1(X)$ so strednou hodnotou $E_1(X)$ a rozdelenie príjmov osôb, ktoré uviedli, že sa necítia byť chudobné ($y_i = 0$), je dané distribučnou funkciou $F_0(X)$ so strednou hodnotou $E_0(X)$. Ďalej nech platí $E_1(X) < E_0(X)$. [4]

Za hranicu subjektívnej chudoby z^* budeme považovať takú hodnotu príjmu X, pri ktorej maximalizujeme rozdiel empirických distribučných funkcií (pozri obr. 3), t. j.

$$z^* = x_0 : \max_{x \in [0;\infty)} \left[F_1(x) - F_0(x) \right]. \tag{2}$$



Obrázok 4: Odhad subjektívnej hranice chudoby na základe ECDF
Zdroj: [4]

Predpokladajme, že $F_1(X)$ môžeme na určitom ohraničenom intervale aproximovať polynomickou funkciou n-tého stupňa $g_1(x)$ a $F_0(X)$ funkciou $g_0(x)$. Označme ich rozdiel ako funkciu $g(x) = g_1(x) - g_0(x)$. Potom za predpokladu $\frac{\partial^2 g(x)}{\partial x^2} < 0$ je odhadom hranice chudoby z^* taký kladný koreň x_0 rovnice

$$\frac{\partial g(x)}{\partial x} = 0, \tag{3}$$

pre ktorý platí:

$$x_0 : \max_x \{ g(x) \}. \tag{4}$$

3. Výsledky

Uveďme najskôr rozdelenie odpovedí na otázku o schopnosti splácať zvyčajné výdavky (premenná *HS120*) a rozdelenie domácností vzhľadom na minimálne požadovaný mesačný príjem (premenná *HS130*).

Z údajov v tabuľke 1 je zrejmé, že keby sme za subjektívne chudobné považovali iba domácnosti schopné splácať zvyčajné mesačné výdavky s veľkými ťažkosťami, bolo by ich 11 %; keby sme za subjektívne chudobné považovali domácnosti schopné splácať zvyčajné mesačné výdavky s ťažkosťami alebo veľkými ťažkosťami, bolo by ich 32 % a keby sme za subjektívne chudobné považovali domácnosti schopné splácať zvyčajné mesačné výdavky s veľkými ťažkosťami, ťažkosťami alebo určitými ťažkosťami, za subjektívne chudobných by sme považovali 75 % domácností SR.

Ako bolo naznačené v 2. kapitole, vychádzajúc z otázky o minimálnom príjme možno za subjektívne chudobné považovať domácnosti, ktoré uviedli hodnotu minimálne

požadovaného príjmu vyššiu, ako je ich disponibilný príjem. Podľa údajov EU-SILC 2009 tomu zodpovedá 59 % domácností.

Tabuľka 1: Rozdelenie domácností vzhľadom na premenné HS120 a HS130

Rozdelenie premennej HS120						
	1	**2**	**3**	**4**	**5**	**6**
f_i	11,1 %	20,5 %	43,6 %	20,6 %	3,5 %	0,7 %
F_i	11,1 %	31,6 %	75,2 %	95,8 %	99,3 %	100,0 %

Rozdelenie domácností vzhľadom na premennú HS130	
$y_i \le z_i$	59,1 %
$y_i \le z_{Me}$	95,5 %

Vysvetlivky: f_i *– relatívne početnosti*
F_i *– kumulatívne relatívne početnosti*
y_i *- aktuálny disponibilný príjem i-tej domácnosti,*
z_i *- min. požadovaný príjem i-tej domácnosti,*
y_{Me} *- mediánový min. požadovaný príjem*

Zdroj: vlastné výpočty podľa údajov EU-SILC 2009 [2]

Naším cieľom je ale určiť hodnotu hranice subjektívnej chudoby v peňažnom vyjadrení, a tak obe premenné je potrebné dať do vzťahu k ekvivalentnému disponibilnému príjmu domácnosti.

S využitím vyššie opísaných postupov (t. j. logit, krivky senzitivity a špecifickosti a porovnanie distribučných funkcií rozdelení príjmov) dostávame nasledovné výsledky:

Tabuľka 2: Odhady hraníc subjektívnej chudoby (EUR) a podielu chudobných domácností

	HS120[+]			**HS130[++]**
	Transf. 1	*Transf. 2*	*Transf. 3*	
logit	585 *(0,4 %)*	3 377 *(11,5 %)*	10 611 *(92,3 %)*	6 735 *(69,2 %)*
senz./špec.	4 717 *(35,3 %)*	5 131 *(44,9 %)*	5 930 *(58,6 %)*	5 568 *(53,0 %)*
distr. fcie	4 874 *(39,0 %)*	5 108 *(44,3 %)*	6 504 *(66,5 %)*	5 680 *(54,8 %)*

[+]*HS120: schopnosť splácať zvyčajné výdavky*
[++]*HS130: minimálne požadovaný príjem*

Zdroj: vlastné výpočty podľa údajov EU-SILC 2009 [2]

S využitím opísaných metód sme získali dvanásť rôznych hraníc subjektívnej chudoby. Na identifikovanie „najlepšej“ hranice možno použiť napríklad porovnanie chyby predpovedí jednotlivých prístupov, no výsledky nie sú jednoznačné.

Odhadnutým hraniciam zodpovedá podiel osôb ohrozených rizikom chudoby v intervale 0,4 % až 92,3 %. Tieto hraničné hodnoty boli získané s použitím logistickej regresie. Takéto nízke/vysoké hodnoty boli pravdepodobne spôsobené nízkym počtom pozorovaní subjektívne chudobných (pri *transf. 1*) resp. nízkym počtom pozorovaní subjektívne nechudobných (pri *transf. 2*). Odhadnuté hodnoty sú výrazne vyššie ako oficiálne publikované hodnoty podielu osôb ohrozených rizikom chudoby ([12], [13], [14]).

4. Záver

Cieľom príspevku bolo popísať a na údajoch EU-SILC 2009 za Slovenskú republiku demonštrovať použitie rôznych metód odhadu peňažných hraníc subjektívnej chudoby domácností. Uvažujúc tri rôzne štatistické metódy odhadu a dve rôzne premenné operacionalizujúce konštrukt subjektívnej chudoby (pričom v prípade jednej z nich prichádzajú do úvahy tri transformácie), dostávame dvanásť rôznych peňažných hraníc subjektívnej chudoby. Možno si preto položiť otázku, či je vôbec možné nájsť tú „správnu" peňažnú hranicu subjektívnej chudoby.

5. Literatúra

[1] ČAPLÁNOVÁ, A. – ORVISKÁ, M..: Determinanty subjektívneho blahobytu jednotlivcov (Dezagregovaná analýza). In: *Ekonomický časopis*. Roč. 58, č. 1 (2010), s. 61 – 73.

[2] ŠÚ SR: *EU SILC 2009, UDB verzia 26/07/2010.* [databáza s mikroúdajmi]. Bratislava: Štatistický úrad SR, 2010.

[3] HUDEC, O. – ŽELINSKÝ, T.: Regionálne disparity na Slovensku hodnotené súhrnným indikátorom úrovne chudoby. In: *Forum Statisticum Slovacum*. Roč. 4, č. 5 (2008). s. 28 – 33.

[4] ŽELINSKÝ, T. – HUDEC, O.: Odhad subjektívnej chudoby na Slovensku založený na distribučnej funkcii rozdelenia príjmov. In: *Forum Statisticum Slovacum*. Roč. 4, č. 7 (2008), s. 152 – 157.

[5] ŽELINSKÝ, T.: Využitie logistickej regresie pri odhadovaní hraníc subjektívnej chudoby. In: *Forum Statisticum Slovacum*. Roč. 5, č. 1 (2009), s. 115 – 119.

[6] ŽELINSKÝ, T.: Pravdepodobnostný prístup k odhadu hraníc subjektívnej chudoby. In: *Slovenská štatistika a demografia*. Roč. 19, č. 1 – 2 (2009). s. 30 – 41.

[7] STANKOVIČOVÁ – I., BARTOŠOVÁ, J.: Príspevok k analýze subjektívnej chudoby v SR a ČR. In: *Forum Statisticum Slovacum*. Roč. 5, č. 3 (2009). s. 151 – 161.

[8] ROCHOVSKÁ, A. – HORŇÁK, M.: Chudoba a jej percepcia v marginálnych regiónoch Slovenska. In: *Geografia Cassoviensis*. Roč. 2, č. 1 (2008). s. 152 – 156.

[9] DUCLOS, J. Y. – ARAAR, A.: *Poverty and Equity: Measurement, Policy and Estimation with DAD*. New York: Springer, 2006. ISBN 978-0387-33318-2.

[10] STANKOVIČOVÁ, I.: Odhad parametrov modelu logistickej regresie metódou maximálnej vierohodnosti. In: *Forum Statisticum Slovacum*. Roč. 4, č. 2 (2008), s. 117 - 128.

[11] HOSMER, D. W. – LEMESHOW, S.: *Applied Logistic Regression*. 2. vyd. New York: John Wiley and Sons, 2000. ISBN 0-471-35632-8.

[12] BARTOŠOVÁ, J. – STANKOVIČOVÁ, I.: *Diferenciace příjmů a chudoba českých a slovenských domácností*. In: Mezinárodní statisticko-ekonomické dny na VŠE v Praze. Praha: VŠE, 2009. ISBN 978-80-86175-66-9.

[13] LABUDOVÁ, V. – SIPKOVÁ, Ľ. *Facts of poverty in Slovakia*. In: Statistics in management of social and economic development. In: 14th Ukrainian-Polish-Slovak scientific conference. Odesa: Palmira, 2008, s. 40 - 48.

[14] ŽELINSKÝ, T.: Analýza chudoby na Slovensku založená na koncepte relatívnej deprivácie. In: *Politická ekonomie*. Roč. 58, č. 4 (2010), s. 542 - 565.

Adresa autora:

Tomáš Želinský, Ing. PhD.,
Ekonomická fakulta, TU Košice, Němcovej 32, 040 01 Košice
e-mail: tomas.zelinsky@tuke.sk

Príspevok bol vytvorený s podporou vedeckovýskumného projektu VEGA 1/0370/08 „*Regionálny prístup k meraniu sociálneho kapitálu a chudoby*" a projektu GAČR 402/09/0515 "*Analýza a modelování finančního potenciálu českých (slovenských) domácností*".

Vplyv indexačnej klauzule
The Impact of the Index Clause

Pavel Zimmermann

Abstract: In non-Life insurance, some bodily liability losses such as loss of income or care costs are paid out regularly as an **annuity** in valorized payments until a specified age or death of the victim. In case of excess of loss reinsurance program, reinsurers often do not accept the inflation risk embedded in such losses and therefore include in the reinsurance treaty the so called **index clause**. The index clause increases the original priority of the reinsurance by the impact of the inflation on the loss. In this article, the formula to calculate the reinsurer's share in case of presence of the index clause is derived and illustrated on numerical example.

Key words: Reinsurance, Annuities, Index clause.

JEL classification: G22

1. Introduction

In non-Life insurance, some bodily liability losses such as loss of income or care costs are paid out regularly as an **annuity** in valorized payments until a specified age or death of the victim. In case of excess of loss reinsurance program, reinsurers often do not accept the inflation risk embedded in such losses and therefore include in the reinsurance treaty the so called **index clause**. The index clause is defined in the actuarial glossary [1] as:

'A clause in an excess of loss reinsurance agreement which shares the effect of future inflation between the insurer and its reinsurers by adjusting the retention and limits in accordance with the movement of a designated index.'

The index clause allows the reinsurer to increase the originally agreed priority by a coefficient which is, roughly said, calculated as a ratio of the sum of all nominal payments to sum of all real (i.e. deflated) payments. (Notice that the reinsurer's share is calculated at the time of the claim closure and therefore past inflation as well as all payments are known.) Discounting is not considered in the priority adjustment.

To the author's knowledge, an analysis of the impact of the insurance clause based on cash flow projection was not published previously. Several notes on the impact of the inflation on the nonproportional reinsurance as well as reasoning for the index clause and reinsurance pricing formulas can be found in [3].

2. Basic Notation and Assumptions

Throughout the article we will denote random variables with capital letters and its values with corresponding small letters. In our calculations, we will first assume only one victim in the age x with known (i.e. deterministic) constant yearly real payment, paid at the beginning of the periods (i.e. annuity in advance), denoted as c. Potential future random changes of the real payment (the 'revision risk') are not considered. Under these assumptions, the sum of the future real payments (i.e. future deflated payments) for a victim at age x with (random) curtate lifetime T_x will be denoted as A'_{T_x}. It can be calculated simply as

$$A'_{T_x} \equiv \sum_{j=0}^{T_x} c = c\left(T_x + 1\right) \tag{1}$$

The index x will be dropped where possible for simplicity. In our calculations, the residual lifetime is in fact the only random variable.

We will also consider some appropriate deterministic future inflation index g (such as wage inflation) by which will the real payments be valorized. For simplicity, the index g will be assumed constant in time. (Further generalization is straightforward.) That means that the sequence of the nominal payments can be calculated as

$$c(1+g)^0, c(1+g)^1, \ldots, c(1+g)^T, \tag{2}$$

The sum of the nominal payments will be denoted A_T, i.e.

$$A_T \equiv \sum_{j=0}^{T} c(1+g)^j = c\frac{1-(1+g)^{T+1}}{1-(1+g)}, \tag{3}$$

which can be further simplified, denoting

$$U_T \equiv \frac{1-(1+g)^{T+1}}{1-(1+g)}, \tag{4}$$

as

$$A_T = c\,U_T. \tag{5}$$

We will assume the excess of loss reisurance treaty with the 'original' (=unadjusted) priority denoted as r'. The reinsurer's share, denoted as Z_T is than calculated as

$$Z_T = \left(A_T - R_T\right)_+ \tag{6}$$

where $\left(f(y)\right)_+ \equiv \max(f(y),0)$ and R_T is the adjusted priority, i.e.

$$R_T = \frac{A_T}{A'_T} r'. \tag{7}$$

3. Expected Reinsurer's Share

For the purposes of reserving or solvency calculation, it is needed to calculate the expected discounted value of the future cash flow of the insurance company (the so called netto best estimate). This is usually performed calculating the expected cash flow without considering the reinsurer's share (the gross best estimate) and then deducting the expected reinsurer's share. The present value of the annuity, i.e. the gross loss will be denoted as $\bar{A}_T$, i.e.

$$\bar{A}_T \equiv \sum_{j=0}^{T} c(1+g)^j v^j \tag{8}$$

where v is the assumed (deterministic) discounting factor. The gross discounted best estimate $E(\bar{A}_T)$ of the annuity is assessed using standard actuarial techniques such as

$$E(\bar{A}_T) = \sum_{t=0}^{\omega-x} \bar{a}_{t\,t} p_x\, q_{x+t} \tag{9}$$

where ${}_t p_x$ denotes the probability that the victim in the age x will survive other t years and q_{x+t} denotes the probability that the victim in the age $x+t$ will die within one year.

The expected reinsurer's share on such annuity loss can then be calculated as

$$E(v^{T_x} Z_{T_x}) = \sum_{t=0}^{\omega-x} z_t \, v^t \, {}_tp_x \, q_{x+t} = \sum_{t=0}^{\omega-x} (a_t - r_t)_+ v^t \, {}_tp_x \, q_{x+t} =$$

$$= \sum_{t=0}^{\omega-x} \left(a_t - \frac{a_t}{a'_t} r' \right)_+ v^t \, {}_tp_x \, q_{x+t} = \sum_{t=0}^{\omega-x} \left(c\,u_t - \frac{c\,u_t}{c\,(t+1)} r' \right)_+ v^t \, {}_tp_x \, q_{x+t},$$

from which

$$E(v^{T_x} Z_{T_x}) = \sum_{t=0}^{\omega-x} \left(c\,u_t \left(1 - \frac{r'}{c\,(t+1)} \right) \right)_+ v^t \, {}_tp_x \, q_{x+t}. \tag{10}$$

Notice that this estimate assumes that the insurance company will receive the reinsurer's share all at once at the claim closure which is somewhat conservative. In reality, there might be some cash calls or deposits received on beforehand. These are however not regular and therefore difficult to capture by a model.

4. Several Victims

So far, only one victim in an age x was assumed. In reality, large bodily losses concern often several victims. In this section, we will consider n victims with vector of ages $\vec{x} \equiv [x_1, x_2, ..., x_n]$ with vector of yearly compensations $\vec{c} \equiv [c_1, c_2, ..., c_n]$. For the following calculations, the joint survival and death probability will be necessary. We will denote the vector of residual lifetimes $\vec{T}_x \equiv [T_{x_1}, T_{x_2}, ..., T_{x_n}]$ and its corresponding values as $\vec{t} \equiv [t_1, t_2, ..., t_n]$. The symbol ${}_{\vec{t}}p_{\vec{x}}$ will denote joint probability that the victim in age x_1 will survive other t_1 years, the victim in age x_2 will survive other t_2 years, ..., and the victim in age x_n will survive other t_n years. Similarly the symbol q_{x+t} will denote the joint probability that the victim in the age $x_1 + t_1$ will die within one year, the victim in the age $x_2 + t_2$ will die within one year, ..., the victim in the age $x_n + t_n$ will die within one year. These probabilities will need to be determined. If no information about the dependency of the residual lifetimes is available, the assumption of independence may be used. (Notice however that this assumption is only an approximation since the victims can be e.g. members of a family etc.) In this case the joint probabilities will be calculated as

$${}_{\vec{t}}p_{\vec{x}} \approx \prod_{k=1}^{n} {}_{t_k}p_{x_k} \tag{11}$$

and

$$q_{\vec{x}+\vec{t}} \approx \prod_{k=1}^{n} q_{x_k+t_k}. \tag{12}$$

Following analogical approach as in case of one victim, we can calculate the expected discounted reinsurer's share as

$$E(v^{\max(\vec{T}_x)} Z_{\vec{T}_x}) =$$

$$= \sum_{t_1=0}^{\omega-x_1} \sum_{t_2=0}^{\omega-x_2} \cdots \sum_{t_n=0}^{\omega-x_n} \left(\sum_{k=1}^{n} c_k \, u_{t_k} - \frac{\sum_{k=1}^{n} c_k \, u_{t_k}}{\sum_{k=1}^{n} c_k \, (t_k+1)} r' \right)_+ v^{\max(\vec{t})} \, {}_{\vec{t}}p_{\vec{x}} \, q_{\vec{x}+\vec{t}} =$$

$$= \sum_{t_1=0}^{\omega-x_1} \sum_{t_2=0}^{\omega-x_2} \cdots \sum_{t_n=0}^{\omega-x_n} \left(\left(\sum_{k=1}^{n} c_k \, u_{t_k} \right) \left(1 - \frac{r'}{\sum_{k=1}^{n} (t_k+1) c_k} \right) \right)_+ v^{\max(\vec{t})} \, {}_{\vec{t}}p_{\vec{x}} \, q_{\vec{x}+\vec{t}},$$

where $v^{\max(\vec{t})}$ denotes $\max_k(t_k)$, $k = 1,\ldots,n$. Notice again, that this estimate assumes that insurance company will receive the reinsurer's share all at once at the claim closure which is in this case the death of the last victim ($\max_k(t_k)$). This can be further simplified using the vector notation $\vec{u} \equiv [u_1, u_2, \ldots, u_k]$ as

$$E(v^{\max(\vec{T}_x)} Z_{\vec{T}_x}) = \sum_{t_1=0}^{\omega-x_1} \sum_{t_2=0}^{\omega-x_2} \ldots \sum_{t_n=0}^{\omega-x_n} \left(\vec{c}\,\vec{u}^{\bullet} \left(1 - \frac{r'}{\vec{t}\,\vec{c}^{\bullet}} \right) \right)_+ v^{\max(\vec{t})}\, {}_{\vec{t}}p_{\vec{x}}\, q_{\vec{x}+\vec{t}}, \tag{13}$$

where $\bullet$ denotes the transposition. Notice again, that this estimate assumes that the insurance company will receive the reinsurer's share all at once at the claim closure which is in this case the death of the last victim ($\max_k(t_k)$).

5. Numerical Illustration

To illustrate the above calculations, three scenarios were analyzed. For each scenario, the following was assumed:

- Only one victim is involved in the loss.
- For simplicity, there is no lump sum payment additional to the annuity.
- The victim is a 30 years old male born on 1.1.1981.
- Generation life tables constructed in [2] updated for the values of the year 2007 were used.
- The original priority $r' =$ CZK 20 000 000.
- Interest rate of 3.5 % ($v = 1/1.035$).
- Yearly valorization of 4%.

Yearly compensations (in CZK) for the three scenarios can be found in the Table 1

Table 1. Yearly compensations (in CZK) for the scenarios considered.

Scenario	Yearly Compensation c
1	250 000
2	500 000
3	750 000

The best estimate of the brutto and netto reserve as well as the reinsurer's share for scenario 1 can be found in Table 2. In this scenario, in the case the index clause is applied, no reinsurance recoveries can be expected - the expected reinsurer's share is zero.

Table 2. The best estimate of the brutto and netto reserve as well as the reinsurer's share for scenario 1.

Scenario 1	Yearly compensation c	Best Estimate		
		Brutto $E(\bar{A}_{Tx})$	Reinsurer's share $E(v^{Tx}Z_{Tx})$	Netto $E(\bar{A}_{Tx})$-$E(v^{Tx}Z_{Tx})$
No Index clause	250 000	15 023 424	3 830 418	11 193 006
Index Clause			0	15 023 424

The results for scenario 2 can be found in Table 3. Here we can observe, that the reinsurer's share in the case the index clause is applied is only about one third of the reinsurer's share where no index clause is applied.

Table 3. The best estimate of the brutto and netto reserve as well as the reinsurer's share for scenario 2.

Scenario 2	Yearly compensation c	Best Estimate		
		Brutto $E(\bar{A}_{Tx})$	Reinsurer's share $E(v^{Tx}Z_{Tx})$	Netto $E(\bar{A}_{Tx})-E(v^{Tx}Z_{Tx})$
No Index clause	500 000	30 046 849	10 668 802	19 378 046
Index Clause			3 719 968	26 326 881

In the scenario 3 (results are displayed in Table 4), the reinsurer's share in the case the index clause is applied is only sixty per cent of the reinsurer's share where the index clause is not applied.

Table 4. The best estimate of the brutto and netto reserve as well as the reinsurer's share for scenario 3.

Scenario 3	Yearly compensation c	Best Estimate		
		Brutto $E(\bar{A}_{Tx})$	Reinsurer's share $E(v^{Tx}Z_{Tx})$	Netto $E(\bar{A}_{Tx})-E(v^{Tx}Z_{Tx})$
No Index clause	750 000	45 070 273	17 728 833	27 341 440
Index Clause			10 433 065	34 637 208

6. Conclusions

In this article, the impact of the index clause on the reinsurer's share was studied. A correct method to calculate the netto best estimate in the case of presence of the index clause in the reinsurance treaty was set up. The impact of the index clause was mapped for different amounts of the yearly payment. For other inputs, realistic (expected) values were used. Due to high amount of the input parameters, general statements for the impact of the index clause are difficult to assess. The impact will always be dependent on the specific values of the parameters. It is apparent form the numerical example that the impact can be massive and a adjustment of the estimates for the index clause should always be performed. Inclusion of the index clause in the calculations of the insurance company can have a dramatic impact on its view on its reserve adequacy, reinsurance program profitability or product profitability.

7. References

[1]AH&T insurance glossary, 2007. Available at *http://www.ahtins.com/glossary/iii/i043.htm*.

[2]CIPRA, T. 1998. Cohort life tables for pension insurance and pension funds in Czech Republic (in Czech), In: Pojistné rozpravy. 1998, Vol. 3, p. 31-57.

[3]FERGUSON, R. E.1974. Nonproportional Reinsurance and the Index Clause, In: CAS Proceedings. 1974, p. 141-164. .

Adresa autora:

Ing. Pavel Zimmermann, Ph.D.
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
zimmerp@vse.cz

Empirical comparison of robustness and efficiency for divergence estimators
Empirické porovnání robustnosti a eficience divergenčních odhadů

Radim Demut, Václav Kůs

Abstract: We introduce minimum Rényi pseudodistance estimators employing adjusting parameter α. We study and compare some general properties of these estimators for different α such as influence curves, robustness and empirical efficiency. We mainly focus on the normal location and scale parameter estimation. In the second part we numerically compare robust properties of minimum Rényi pseudodistance estimators for different parameters α with minimum Hellinger distance estimators and some other estimators. This comparisons are performed within the normal family under different level of contaminations.

Keywords: Minimum distance estimation, Robustness, Empirical efficiency, PC simulation.
Klíčová slova: Odhady s minimální vzdáleností, robustnost, empirická eficience, PCsimulace.
JEL classification: C13

1. Rényi pseudodistance estimators

In this section we are interested in pseudodistances, which have been recently introduced in [1]. They are not classical distances because the symmetry or the triangle inequality need not be valid. Let $\mathsf{P}=\{P_\theta:\theta\in\Theta\subset\mathsf{R}^m\}$ be a set of probability measures on measurable space (X,A). Our estimates will be based on the identically independently distributed observations $X_1,\dots,X_n$ governed by P_r. Since we are interested in robustness, we allow the case when $P_r\notin\mathsf{P}$. Therefore we introduce another set $\mathsf{P}^+=\mathsf{P}\cup\{P_r\}$. Moreover, we assume that every probability measure $Q\in\mathsf{P}^+$ is dominated by a σ-finite measure λ with the density $q=\frac{\mathrm{d}Q}{\mathrm{d}\lambda}$, for all $Q\in\mathsf{P}^+$. Further we assume, that for all $P,Q\in\mathsf{P}^+$ it holds that $pq>0,\lambda-\text{a.e.}$, where p,q are the densities corresponding to P,Q. We denote the set of all empirical distribution functions P_{emp}. The following theorem correctly defines these pseudodistances.

Theorem 1. Let for some $\beta>0$ it hold that $p^\beta,q^\beta,\ln p\in\mathrm{L}_1(Q)$, for all $P\in\mathsf{P},Q\in\mathsf{P}^+$. Then for all α, $0<\alpha\le\beta$, and for $P\in\mathsf{P},Q\in\mathsf{P}^+$ the following expression

$$\mathsf{R}_\alpha(P,Q)=\frac{1}{1+\alpha}\ln\Big(\int p^\alpha\mathrm{d}P\Big)+\frac{1}{\alpha(1+\alpha)}\ln\Big(\int q^\alpha\mathrm{d}Q\Big)-\frac{1}{\alpha}\ln\Big(\int p^\alpha\mathrm{d}Q\Big)$$

represents the family of decomposable pseudodistances

$$\mathsf{R}_\alpha(P,Q)=\mathsf{R}^0_\alpha(P)+\mathsf{R}^1_\alpha(Q)-\frac{1}{\alpha}\ln\Big(\int p^\alpha\mathrm{d}Q\Big),$$

$$\mathsf{R}^0_\alpha(P)=\frac{1}{1+\alpha}\ln\Big(\int p^\alpha\mathrm{d}P\Big),\qquad \mathsf{R}^1_\alpha(Q)=\frac{1}{\alpha(1+\alpha)}\ln\Big(\int q^\alpha\mathrm{d}Q\Big).$$

Moreover, for $\alpha\to0$ it holds

$$\mathsf{R}_0(P,Q) = \lim_{\alpha\to 0}\mathsf{R}_\alpha(P,Q) = \int(\ln q - \ln p)\mathrm{d}Q.$$

We are interested in the estimators obtained by replacing the hypothetical distribution P_r in the pseudodistances $\mathsf{R}_\alpha(P_\theta, P_r)$ by the empirical distribution P_n. That means we are interested in the family of minimum Rényi pseudodistance estimators defined as $\theta_{n,\alpha} = T_\alpha(P_n)$ for $T_\alpha(Q) \in \Theta$ with $Q \in \mathsf{P}^+ \cup \mathsf{P}_{\mathrm{emp}}$ satisfying the condition

$$T_\alpha(Q) = \begin{cases} \operatorname{argmin}_\theta[\frac{1}{1+\alpha}\ln(\int p_\theta^\alpha \mathrm{d}P_\theta) - \frac{1}{\alpha}\ln(\int p_\theta^\alpha \mathrm{d}Q)] & \text{for} 0 < \alpha \le \beta, \\ \operatorname{argmin}_\theta - \ln(\int p_\theta \mathrm{d}Q) & \text{for} \alpha = 0. \end{cases}$$

If we replace Q by the empirical distribution function P_n we get the estimator

$$\theta_{\alpha,n} = \begin{cases} \operatorname{argmax}_\theta C_\alpha(\theta)^{-1}\frac{1}{n}\sum_{i=1}^n p_\theta^\alpha(X_i) & \text{for} 0 < \alpha \le \beta, \\ \operatorname{argmax}_\theta \frac{1}{n}\sum_{i=1}^n \ln p_\theta(X_i) & \text{for} \alpha = 0, \end{cases}$$

where $C_\alpha(\theta) = (\int p_\theta^{1+\alpha}\mathrm{d}\lambda)^{\alpha/1+\alpha}$. If for a fixed n the argument maxima is in a compact subset of Θ and the MLE $\theta_{0,n}$ is unique, than $\lim_{\alpha\to 0}\theta_{\alpha,n} = \theta_{0,n}$.

2. Application to the normal family

We use min R_α-estimators for the mean and variance in the normal family, i.e. $\theta = (\mu, \sigma)$. For $\alpha = 0$ the estimator coincides with MLE

$$\theta_{\mathsf{R}_0,n} = \operatorname{argmax}_\theta \frac{1}{n}\sum_{i=1}^n \ln[\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(X_i-\mu)^2}{2\sigma^2})].$$

For $\alpha > 0$, the constant $C_\alpha(\theta)$ can be computed as $\int p_\theta^{1+\alpha}(x)\mathrm{d}x = [(1+\alpha)^{1/2}(2\pi\sigma^2)^{\alpha/2}]^{-1}$, which gives us the min R_α-estimator

$$\theta_{\mathsf{R}_\alpha,n} = \operatorname{argmax}_\theta \frac{1}{n\sigma^{\alpha/(1+\alpha)}}\sum_{i=1}^n \exp(-\alpha\frac{(X_i-\mu)^2}{2\sigma^2}).$$

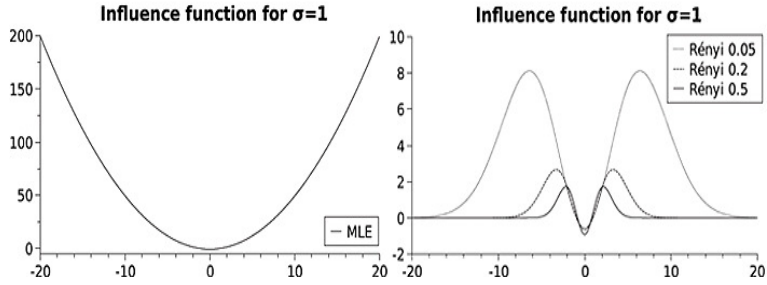


Figure 1: Min R_α--estimator IF for $\alpha = 0$ (MLE) and $\alpha = 0.05, 0.2, 0.5$.

The influence functions for minimum Rényi estimators are computed in [1],

$$\mathrm{IF}(x;T_{\mathsf{R}_\alpha},\sigma)=\frac{(1+\alpha)^{5/2}\sigma}{2}[((\frac{x}{\sigma})^2-\frac{1}{1+\alpha})\exp(-\frac{\alpha x^2}{2\sigma^2})],$$

becoming the influence function of MLE for $\alpha=0$. The minimum Rényi estimators are robust for $\alpha>0$ in the sense that their influence functions are bounded. Moreover, for $\alpha>0$, it holds that $\lim_{|x|\to\infty}\mathrm{IF}(x;T_{\mathsf{R}_\alpha},\sigma)=0$, see Figure 1.

We have to overcome a singularity in maximization of the R_α pseudodistances for $\alpha>0$ in the normal family. If we choose $\mu=X_i$ for any $i\in\{1,\dots,n\}$, it holds

$$\lim_{\sigma\downarrow 0}\frac{1}{n\sigma^{\alpha/(1+\alpha)}}\sum_{i=1}^{n}\exp(-\alpha\frac{(X_i-\mu)^2}{2\sigma^2})=+\infty$$

and thus we cannot find the global maximum of the function. We solve this problem by the following trick. Instead of minimizing over (μ,σ) in the set $(-\infty,+\infty)\times(0,+\infty)$ we restrict the region to the smaller set $(-\infty,+\infty)\times(\delta,+\infty)$ for some $\delta>0$ to avoid the singularity.

3. Computer simulations and comparisons

In this chapter we present results of our simulation. We compare minimum Rényi pseudodistance estimators for different α with minimum Hellinger distance estimator, minimum LeCam divergence estimator and minimum Kolmogorov estimator. We present results for the mean μ and scale parameter σ in normal family. Finally, we apply our robust estimates to the real data cited in [2].

Let us assume we have the normal distribution $N(0,1)$ contaminated by another normal distribution $N(0,100)$ with level of contamination ε. We generate hundred observations from this distribution and estimate the mean and scale in the normal family. This procedure is repeated thousand times. Then we tabulate the mean values of the estimates (denoted by $m(\mu)$ and $m(\sigma)$) and the variance of the estimates (denoted by $s(\mu)$ and $s(\sigma)$) and also the relative efficiency of the estimates of parameter σ defined by

$$\mathrm{eef}(\sigma)=\sqrt{\frac{\frac{1}{m}\sum_{k=1}^{m}(\hat{\sigma}_{0,k}-\sigma_{\mathrm{real}})^2}{\frac{1}{m}\sum_{k=1}^{m}(\hat{\sigma}_{k}-\sigma_{\mathrm{real}})^2}}=\sqrt{\frac{\frac{1}{m}\sum_{k=1}^{m}(\hat{\sigma}_{0,k}-1)^2}{\frac{1}{m}\sum_{k=1}^{m}(\hat{\sigma}_{k}-1)^2}},$$

where $m=1000$ is the number of repetitions, $\hat{\sigma}_{0,k}$ stands for the maximum likelihood estimator of σ and $\hat{\sigma}_k$ denote our minimum distance estimators for $k=1,\dots,m$.

In Figure 2 we present rectangular graphs, where the black dot is the desired true value and the rectangle is kind of empirical confidence region, on the horizontal axis is the interval $(m(\mu)-s(\mu),m(\mu)+s(\mu))$ and on the vertical axis is the interval $(m(\sigma)-s(\sigma),m(\sigma)+s(\sigma))$. We can see from Figure 2 that for increasing α the robustness of min R_α-estimator also increases. Even for $\alpha=0.05$ the min R_α-estimator is much better than MLE for small contamination.

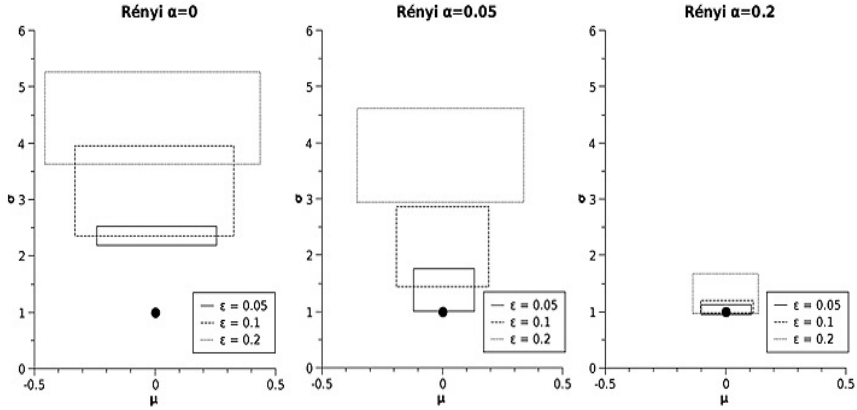


Figure 2: Comparison of robust properties of min R_α-estimators under different α and various ε contaminations.

Figure 3 shows the comparison of min R_α-estimator for $\alpha = 0.2, 0.5$ with minimum Hellinger distance estimator (denoted by Phi1) using the Hellinger distance

$$H(F,G) = \left(2\left(1 - \int \sqrt{fg}\,\mathrm{d}\mu\right)\right)^{\frac{1}{2}} = \left(\int (\sqrt{f} - \sqrt{g})^2\,\mathrm{d}\mu\right)^{\frac{1}{2}},$$

and with minimum LeCam divergence (denoted by Phi2) defined by means of

$$LC(F,G) = \int \frac{(f-g)^2}{f+g}\,\mathrm{d}\mu,$$

where f, g denote densities corresponding to the distributions F, G. Here, in both these divergences H and LC, we used the kernel estimates with normal kernel instead of the empirical estimates of the densities. Finally, we compare these estimators with the minimum Kolmogorov estimator. The most robust estimator is min R_α-estimator for $\alpha = 0.5$, unfortunately, it is not so good in the non-contaminated model. The second best estimator for small contaminations along with very good performances for big contaminations is the min R_α-estimator for $\alpha = 0.2$. This estimator is also the best in the non-contaminated model. The minimum Kolmogorov estimator does not perform too well how can be seen from Figure 3.

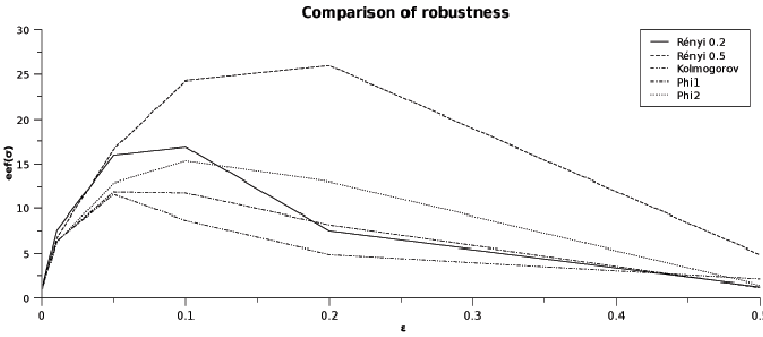


Figure 3: Comparison of relative efficiency of σ estimate for min R_α, $\alpha = 0.2, 0.5$, Hellinger, Kolmogorov, LeCam estimators under contaminations ε.

4. Speed of light Newcomb's data

Finally we present our results for the Newcomb's data of the speed of light measurement. These data are presented in Table 1 taken over from [2]. If we analyze these data and apply the normal model it can be seen that the numbers -2 and -44 are apparent outliers. So we try to compute our estimates for the purified data, that means, we remove numbers -2 and -44 from the data set. This estimate is in the „correct" column of Table 2. In the next column of this table the estimates of the cleared data with added -2 are given. Then we added -2 and 12, then -2 and -44 and finally -2 and -100 to the cleared data. Moreover, $\mu_{R,\alpha}$ and $\sigma_{R,\alpha}$ denote min R_α estimate of the mean and scale, respectively.

We can see from Table 2, that the min R_α estimate for $\alpha = 0.1$ is not too far away from the MLE for $\alpha = 0$ in the „correct" column, but it is much better for the contaminated data. Also the min R_α estimate for $\alpha = 0.2$ is very good in the „correct" case. Moreover, it also performs really well with added contaminations. For $\alpha = 0.5$ the estimator is even more robust but it does not perform so well in the „correct" case.

28	26	33	24	34	27	16	40	-2	29	22	24
21	25	30	23	29	31	19	24	20	36	32	36
28	25	21	28	29	37	25	28	26	30	32	36
26	30	22	36	23	27	27	28	27	31	27	26
33	26	32	32	24	39	28	24	25	32	25	29
27	28	29	16	23	-44						

Table 1: Newcomb's data of the speed of light measurement.

$\alpha \backslash n$	„Correct“	-2	$-2{,}12$	$-2{,}2$	$-2{,}{-}44$	$-2{,}{-}100$
	$\mu_{R,\alpha}$	$\mu_{R,\alpha}$	$\mu_{R,\alpha}$	$\mu_{R,\alpha}$	$\mu_{R,\alpha}$	$\mu_{R,\alpha}$
	$\sigma_{R,\alpha}$	$\sigma_{R,\alpha}$	$\sigma_{R,\alpha}$	$\sigma_{R,\alpha}$	$\sigma_{R,\alpha}$	$\sigma_{R,\alpha}$
0	27.750	27.292	27.061	26.909	26.212	25.364
	5.044	6.201	6.431	6.886	10.664	16.723
0.01	27.746	27.335	27.103	26.964	26.463	25.978
	5.043	6.108	6.351	6.794	9.888	14.358
0.1	27.710	27.600	27.396	27.401	27.599	27.600
	5.029	5.383	5.695	5.884	5.385	5.383
0.2	27.666	27.649	27.517	27.603	27.649	27.649
	4.999	5.069	5.312	5.220	5.069	5.069
0.5	27.515	27.515	27.487	27.515	27.515	27.515
	4.822	4.822	4.914	4.825	4.822	4.822

Table 2: R_α***-estimator for various contamination of Newcomb's data.***

5. References

[1] BRONIATOWSKI, M. – VAJDA, I. 2009. Several Applications of Divergence Criteria in Continuous Families, Research report No 2257 September 2009, ÚTIA AV Č R, Praha, 2009.

[2] BASU, A. – HARRIS, I.R. – HJORT, N.L. – JONES M.C. 1998. Robust and efficient estimation by minimizing a density power divergence, Biometrika, 85, 549-559, 1998.

[3] DEMUT, R. 2010. *Robustnost odhadu hustot s minimální divergencí*, Diploma Thesis, Czech Technical University, 2010.

[4] FISHMAN, G.S. 1996. Monte Carlo: concepts, algorithms, and application, Springer-Verlag New York, 1996.

Adresa autora (-ov):

Radim Demut, Ing
Katedra matematiky, FJFI ČVUT
Trojanova 13, 120 00 Praha 2
demut@seznam.cz

Václav Kůs, Ing, PhD.
Katedra matematiky, FJFI ČVUT
Trojanova 13, 120 00 Praha 2
vaclav.kus@fjfi.cvut.cz

This work has been supported by the grant SGS OHK4-007/10.

Power subdivergence estimators in the normal model – PC simulation
Power subdivergenční odhady v normálním modelu – PC simulace

Iva Frýdlová, Václav Kůs

Abstract: We present the results of extensive simulation study of minimum divergence based estimators. Broniatowski and Vajda in [2] proposed several modifications of the minimum divergence rule to provide alternative *maximum subdivergence estimators with escort parameter* $\theta \in \Theta$ and *minimum superdivergence estimators*. The main interest of our research was to examine these modifications in practical use as to the consistency, robustness and efficiency of the estimators. We focus on the well known family of power divergences parametrized by $\alpha \in \mathsf{R}$ in the normal distribution model. We run a comparative computer simulation for several randomly selected contaminated and uncontaminated data sets and for different sample sizes and different ϕ-divergence parameters. We shall focus on the subdivergence estimators and their performances.

Key words: Minimum divergence estimation, Simulations, Robustness, Empirical efficiency.

Klíčová slova: Odhady s minimální vzdáleností, simulace, robustnost, empirická efficience.

JEL classification: C13

1. Introduction

As was already mentioned in many publications, the well known information-theoretic measures of divergence of probability measures cannot be directly applied in statistical estimation, since the divergence between the theoretical absolutely continuous probability measure and the discrete empirical probability measure is always equal to its upper bound and often takes on infinite values. In 2008 - 2009 Broniatowski & Vajda ([2]) studied and extended two different modifications of divergences proposed independently in 2006 ([5], [1]). One of them was the concept of *subdivergences* and corresponding new *maximum subdivergence estimator with escort parameter* θ. The main interest of our research ([4]) was to examine the consistency, robustness and efficiency of the estimator with respect to the standard maximum likelihood estimator.

2. ϕ-divergences and minimum ϕ-divergence estimators

Let (X, A) be a measurable space and let P be a set of all probability measures on (X, A). If $P \in \mathsf{P}$ is dominated by a σ-finite measure λ on (X, A), then $p = dP/d\lambda$ is a Radon-Nikodym density of P with respect to measure λ.

Definition 1. Let $P, Q \in \mathsf{P}$, $\{P, Q\} \ll \lambda$, $p = dP/d\lambda$ and $q = dQ/d\lambda$. *ϕ-divergence of distributions P and Q* is a function $D_\phi : \mathsf{P} \times \mathsf{P} \to [0, \infty]$ defined by

$$D_\phi(P, Q) = \int_{\mathsf{X}} \phi\left(\frac{p}{q}\right) dQ = \int_{\mathsf{X}} q\, \phi\left(\frac{p}{q}\right) d\lambda \quad , \tag{1}$$

where $\phi : (0,\infty) \to \mathsf{R}$ is a convex function called a *generating function*. For this formula to be well defined, we denote $\phi(0) := \lim_{t\to 0_+} \phi(t)$ and $\phi(\infty)/\infty := \lim_{t\to\infty} \frac{\phi(t)}{t}$ ("$0\cdot\infty = 0$") and we put

$$q\,\phi\left(\frac{p}{q}\right) = \begin{cases} q\,\phi(0) & \text{if } p = 0 \\ p\,\phi(\infty)/\infty & \text{if } q = 0, \end{cases}$$

For each generating function ϕ it holds $D_\phi(P,Q) \le \phi(0) + \phi(\infty)/\infty$ for every $P, Q \in \mathsf{P}$, and the equality takes place if $P \perp Q$, i.e. P, Q are singular. Let $\mathbf{\Phi}$ be the class of all functions ϕ which are twice differentiable, strictly convex generating functions with $\phi(1) = 0$. Moreover, we deal only with P and Q which are either measure-theoretically equivalent, $P \equiv Q$, or measure-theoretically orthogonal, $P \perp Q$.

In the sequel, we use the power divergences $D_\alpha(P,Q) := D_{\phi_\alpha}(P,Q), \quad \alpha \in \mathsf{R}$, where

$$\phi_\alpha(t) = \frac{t^\alpha - \alpha(t-1) - 1}{\alpha(\alpha-1)} \quad \alpha \neq 0, \alpha \neq 1 \tag{3}$$

with the limiting cases $\phi_0(t) = -\ln t + t - 1$ *and* $\phi_1(t) = t\ln t - t + 1.$ (4)

For these functions it holds $0 \le D_\alpha(P,Q) \le \begin{cases} \frac{1}{\alpha(1-\alpha)} & \text{if} \quad 0 < \alpha < 1 \\ \infty & \text{otherwise.} \end{cases}$ (5)

The left equality takes place if and only if $P = Q$. For $0 < \alpha < 1$, the right equality takes place if and only if $P \perp Q$. Otherwise it takes place if $\alpha \le 0$ and Q is not absolutely continuous w.r.t. P or if $\alpha \ge 1$ and P is not absolutely continuous w.r.t. Q.

Now, let $X_1, X_2, ..., X_n$ be independent and identically distributed observations governed by $P_{\theta_0} \in \mathsf{P}$, where $\mathsf{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measures on (X, A), $\Theta \subset \mathsf{R}^d$, and we assume that for every $\theta, \theta_0 \in \Theta$, $\theta \neq \theta_0$ holds $P_\theta \neq P_{\theta_0}$ and $P_\theta \equiv P_{\theta_0}$. Moreover, we assume the family P to be nonatomic (continuous), i.e. for all $\theta \in \Theta$ and $x \in \mathsf{X}$ we require $P_\theta(\{x\}) = 0$. We also let the data $X_1, X_2, ..., X_n$ to be represented by an empirical probability measure $P_n = \frac{1}{n}\sum_{i=1}^n P_{X_i}$, where P_x is the Dirac probability measure at the point $x \in \mathsf{X}$.

Definition 2. Let $\phi \in \mathbf{\Phi}$. We say that an estimator $\hat{\theta}_n : \mathsf{X}^n \to \Theta$ of the true parameter $\theta_0 \in \Theta$ is a *minimum ϕ-divergence estimator* if $\hat{\theta}_n = \operatorname{argmin}_\theta D_\phi(P_\theta, P_n)$.

The problem we encounter with these estimators is that the continuous family P and the family of empirical distributions P_{emp} are measure-theoretically orthogonal, i.e. $P_\theta \perp P_n$ for every $P_\theta \in \mathsf{P}$ and $P_n \in \mathsf{P}_{emp}$. This implies for every $P_\theta \in \mathsf{P}$ and $P_n \in \mathsf{P}_{emp}$ that $D_\phi(P_\theta, P_n) = \phi_\alpha(0) + \phi_\alpha(\infty)/\infty$ and the above defined estimates are trivial.

To face this problem, it is possible to use some prior smoothing of the data or another nonparametric density estimation like we did by implementing histogram in [3], but these

methods bring another unpleasant obstructions such as bandwidth selection. In the next section, we present several modifications of the minimum divergence rule studied by Broniatowski & Vajda ([2]) avoiding these complications.

3. Power subdivergence estimators

We shall regard the probability measures $P \in \mathsf{P}$ and $Q \in \mathsf{Q}$ for $\mathsf{Q} = \mathsf{P} \cup \mathsf{P}_{emp}$. Measures P, Q are either measure-theoretically equivalent (if $Q \in \mathsf{P}$) or measure-theoretically orthogonal (if $Q \in \mathsf{P}_{emp}$). Consider the family of finite expectations

$$\underset{\sim}{\mathrm{D}}_{\phi,\tilde{\theta}}(P_\theta, Q) = \int \phi'(p_\theta / p_{\tilde{\theta}})\, \mathrm{d}P_\theta + \int \phi^{\#}(p_\theta / p_{\tilde{\theta}})\, \mathrm{d}Q, (P_\theta, Q) \in \mathsf{P} \otimes \mathsf{Q} \tag{6}$$

parametrized by $(\phi, \tilde{\theta}) \in \mathbf{\Phi} \otimes \mathbf{\Theta}$, where $\phi^{\#}(t) = \phi(t) - t\phi'(t)$ *for every* $\phi \in \mathbf{\Phi}$, and ϕ' denotes the derivative of ϕ. For (6) to be correctly defined, we assume that the integrals exist and have a finite value. Now, the *maximum subdivergence estimators* (briefly, the $\max \underset{\sim}{\mathrm{D}}_{\phi}$-estimators) with escort parameter $\theta \in \mathbf{\Theta}$ are defined as

$$\tilde{\theta}_{\phi,\theta,n} = argmax_{\tilde{\theta}}\ \underset{\sim}{D}_{\phi,\tilde{\theta}}(P_\theta, P_n) = argmax_{\tilde{\theta}} \left[\int \phi' \left(\frac{p_\theta}{p_{\tilde{\theta}}} \right) \mathrm{d}P_\theta + \frac{1}{n} \sum_{i=1}^{n} \phi^{\#} \left(\frac{p_\theta(X_i)}{p_{\tilde{\theta}}(X_i)} \right) \right]. \tag{7}$$

If we restrict ourselves to a subclass of these estimators determined by the power divergences (2), and if we consider only a standard normal model $\mathsf{P} = \{P_{\mu,\sigma} : \mu \in \mathsf{R},\ \sigma > 0\}$ with the location parameter μ and scale σ, we receive for $\alpha = 0$

$$(\tilde{\mu}_{0,\mu,\sigma,n}, \tilde{\sigma}_{0,\mu,\sigma,n}) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i, \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \tilde{\mu}_{0,\mu,\sigma,n})^2} \right) \tag{8}$$

which is the maximum likelihood estimate in the family of normal distributions. Hence, the class of $\max \underset{\sim}{\mathrm{D}}_{\phi}$-estimators are extensions of the MLE. For $\alpha > 0$, $\alpha \neq 1$ the estimators become

$$\tilde{\mu}_{\alpha,\mu,\sigma,n} = argmin_{\tilde{\mu},\tilde{\sigma}}\ M_{\alpha,\mu,\sigma}(P_n, \tilde{\mu}, \tilde{\sigma}) \tag{9}$$

with

$$M_{\alpha,\mu,\sigma}(P_n, \tilde{\mu}, \tilde{\sigma}) = \frac{1}{1-\alpha} \int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} + \frac{1}{\alpha n} \sum_{i=1}^{n} \left(\frac{p_{\mu,\sigma}(X_i)}{p_{\tilde{\mu},\tilde{\sigma}}(X_i)} \right)^{\alpha} \tag{10}$$

where

$$\left(\frac{p_{\mu,\sigma}(x)}{p_{\tilde{\mu},\tilde{\sigma}}(x)} \right)^{\alpha} = \left(\frac{\tilde{\sigma}}{\sigma} \right)^{\alpha} \exp\left\{ \frac{\alpha(x-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{\alpha(x-\mu)^2}{2\sigma^2} \right\}, \tag{11}$$

and

$$\int \left(\frac{p_{\mu,\sigma}}{p_{\tilde{\mu},\tilde{\sigma}}} \right)^{\alpha} \mathrm{d}P_{\tilde{\mu},\tilde{\sigma}} = \exp\left\{ \frac{-\alpha(1-\alpha)(\mu-\tilde{\mu})^2}{2[\alpha\tilde{\sigma}^2 + (1-\alpha)\sigma^2]} - \ln \frac{\sqrt{\alpha\tilde{\sigma}^2 + (1-\alpha)\sigma^2}}{\tilde{\sigma}^{\alpha}\sigma^{1-\alpha}} \right\}. \tag{12}$$

For $\alpha = 1$ we get

$$M_{1,\mu,\sigma}(P_n, \tilde{\mu}, \tilde{\sigma}) = \lim_{\alpha \to 1} M_{\alpha,\mu,\sigma}(P_n, \tilde{\mu}, \tilde{\sigma})$$

$$= \frac{-(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{1}{2} \left[-\ln \left(\frac{\sigma}{\tilde{\sigma}} \right)^2 + \left(\frac{\sigma}{\tilde{\sigma}} \right)^2 - 1 \right] + \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tilde{\sigma}}{\sigma} \right) \exp\left\{ \frac{(X_i-\tilde{\mu})^2}{2\tilde{\sigma}^2} - \frac{(X_i-\mu)^2}{2\sigma^2} \right\}. \tag{13}$$

4. Computer Simulations

First we inspect the estimators in normal model of location ($\sigma = 1$ fixed) and then in the normal scale model ($\mu = 0$ fixed). Let $X_1,...,X_n$ be observations from the convex mixture $P_\varepsilon = (1-\varepsilon)P + \varepsilon Q$. Here P is a standard normal model with location $\mu = 0$ and scale $\sigma = 1$, further denoted by $N(0,1)$, and Q is successively normal $(N(0,9), N(0,100))$, logistic $(Lo(0,1))$, and Cauchy $(C(0,1))$ distribution. The contaminations we use are 0, 1, 5, 10, 20, 30 percent respectively, i.e. ε takes on the values $0,0.01,0.05,0.1,0.2$, and 0.3. The sample size n is considered successively 20, 50, 100, 200, 500. In case of $\max \underline{D}_\alpha$-estimators $\tilde{\mu}_{\alpha,\mu,n}$ we take into account power parameters 0, 0.01, 0.05, 0.1, 0.2, and 0.5 and we select the escort parameters $\mu = 0,0.1,0.2,0.5,1$ and finally $\mu = \overline{\mathbf{X}}_n$ (MLE) and $\mu = med_i(X_i)$. For $\max \underline{D}_\alpha$-estimators $\tilde{\sigma}_{\alpha,\sigma,n}$ we consider the same values of power parameter and the escort parameters $\sigma^2 = 0.5,1,1.2,1.5,2$ and finally $\sigma = \mathbf{S}_n$ (MLE) and $\sigma = med_j(|X_j|)$, which is the MAD estimate of scale for known location parameter $\mu = 0$ (short of the calibration constant), otherwise MAD $= 1.483\, med_j(|X_j - med_i(X_i)|)$.

To evaluate the behavior of power superdivergence (or power subdivergence) estimators we generate K different data samples (K=100 for superdivergence estimators, or K=1000 for subdivergence estimators) to gain K different estimates (further indexed by $^{(k)}$). We compute means and standard deviations

$$m(\tilde{\mu}) = \frac{1}{K}\sum_{k=1}^{K}\mu_{\alpha,n}^{(k)} \qquad s(\tilde{\mu}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mu_{\alpha,n}^{(k)} - m(\tilde{\mu}))^2}$$

$$m(\tilde{\sigma}) = \frac{1}{K}\sum_{k=1}^{K}\sigma_{\alpha,n}^{(k)} \qquad s(\tilde{\sigma}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - m(\tilde{\sigma}))^2}$$

of the $\min \overline{D}_\alpha$-estimators (or $\max \underline{D}_\alpha$-estimators) and maximum likelihood estimators $\overline{\mathbf{X}}_n^{(k)}$ and $\mathbf{S}_n^{(k)}$. Making use of these we receive the relative empirical efficiencies

$$\mathrm{eref}(\tilde{\mu}) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\overline{\mathbf{X}}_n^{(k)})^2}{\frac{1}{K}\sum_{k=1}^{K}(\mu_{\alpha,n}^{(k)})^2} \quad \textit{and} \quad \mathrm{eref}(\tilde{\sigma}) = \frac{\frac{1}{K}\sum_{k=1}^{K}(\mathbf{S}_n^{(k)} - 1)^2}{\frac{1}{K}\sum_{k=1}^{K}(\sigma_{\alpha,n}^{(k)} - 1)^2}.$$

Results for power subdivergence estimators of location. For power parameter $\alpha = 0$ we can conclude that the estimates coincide with MLE, i.e. $\mathrm{eref}(\tilde{\mu}) = 1$, as was expected. In case of escort parameter $\mu = 0$, the $\max \underline{D}_\alpha$-estimators for the uncontaminated data still more or less copy the behavior of MLE even for values of $\alpha > 0$, but as the contamination grows, we observe that the means and standard deviations of $\max \underline{D}_\alpha$-estimator move apart from MLE taking on lower values than maximum likelihood estimate of the contaminated data. In case of $m(\tilde{\mu})$ the difference is only slight (yet favourable), but in case of $s(\tilde{\mu})$ the difference is apparent and causes a fair increase in empirical relative efficiency. Especially in the case of contamination by Cauchy distribution (which generates distant outliers) the robustness of the estimator escorted by $\mu = 0$ is rather stunning compared to maximum likelihood estimator.

However, all that was stated above holds only for $\mu = 0$. The situation changes to the worse for the escort parameter μ approaching 1. The consistency, efficiency, even the robustness tendencies slowly vanish for ε increasing, and we see that apart from the case of $\mu = 0$ the $\max \underline{D}_{\alpha}$-estimators do not possess the useful properties we would desire.

In accordance with the fact that the only reasonable results we obtained were for $\mu = 0$ which is the true parameter of the estimated data, some very good results were received for the value of the escort parameter $\mu = \overline{\mathbf{X}}_n$. For contaminations by $N(0,9)$, $N(0,100)$ and $Lo(0,1)$ we received perfect match with MLE for all values of ε. Nevertheless, an outstanding behavior was noticed in case of Cauchy contamination, where the power subdivergence estimator shows a significant resistance to distant outliers. In this situation, the standard deviation of the maximum likelihood estimator with great volatility copies the occurrence of extreme outliers, while the standard deviation of MLE-escorted subdivergence estimator retains low values with steady convergence to 0. This fact results also in huge empirical relative efficiency.

Even though these results are very encouraging, some much better results were obtained for estimators escorted by median, which is known to be a robust estimate. Although, the efficiencies are not as good as for $\mu = 0$, it is clear that the choice of median for escort parameter produces much more robust estimates that those escorted by MLE. For not too high levels of contamination the subdivergence estimators show even higher, better robustness than the original inserted median itself. We show this in Table 1 by comparing subdivergence estimators with median by plugging the median into the *eref* formula instead of MLE, and observing the $eref(\tilde{\mu})$ increasing.

Table 1: Max $\underline{D}_{\alpha}$ ***-estimator: data*** $0.9N(0,1) + 0.1Lo(0,1)$***, escort parameter*** $\mu = med_i(X_i)$

α/n	50			100			200			500		
	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$eref(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$eref(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$eref(\tilde{\mu})$	$m(\tilde{\mu})$	$s(\tilde{\mu})$	$eref(\tilde{\mu})$
MED	−0.006	0.185	1.000	0.003	0.130	1.000	−0.001	0.093	1.000	−0.002	0.059	1.000
0.00	−0.003	0.161	1.323	0.004	0.112	1.348	−0.000	0.079	1.395	−0.001	0.050	1.417
0.01	−0.003	0.161	1.324	0.004	0.112	1.348	−0.000	0.079	1.396	−0.001	0.050	1.418
0.05	−0.003	0.161	1.330	0.004	0.112	1.352	−0.000	0.078	1.398	−0.001	0.050	1.419
0.10	−0.003	0.160	1.336	0.004	0.112	1.355	−0.000	0.078	1.401	−0.001	0.050	1.421
0.20	−0.003	0.160	1.345	0.003	0.111	1.360	−0.000	0.078	1.405	−0.001	0.050	1.424
0.50	−0.003	0.159	1.357	0.003	0.111	1.365	−0.000	0.078	1.409	−0.001	0.050	1.427

Results for power subdivergence estimators of scale. As expected, for $\alpha = 0$ we get the exact MLE, hence $\mathrm{eref}(\tilde{\sigma})$ is always equal to 1. For $\varepsilon = 0$, i.e. the uncontaminated data, the subdivergence estimators more or less correspond with the maximum likelihood estimators,

but they do not outperform them. With rising value of parameter α also the standard deviation $s(\tilde{\sigma})$ rises a little, which causes a certain loss of efficiency. For $\varepsilon > 0$ and escort parameters $\sigma^2 > 1$ we observe a loss of consistency, however the MLE loses its consistency too, and with rising contamination we see that the $\max \underline{D}_\alpha$-estimators possess lower values of means and standard deviations then MLE. The best results were received mostly for escort parameter $\sigma^2 \in (0.5,1\rangle$. Here, the estimates retained the consistency even for highly contaminated data, showed substantially lower values of $m(\tilde{\sigma})$ and $s(\tilde{\sigma})$, which resulted in high empirical relative efficiency (cf. Table 2). For data contaminated by $N(0,100)$ and $C(0,1)$, the subdivergence estimators perform better then MLE even for very small level of contamination $\varepsilon = 0.01$ and all values of escort parameter σ.

Table 2: Max $\underline{D}_\alpha$ ***-estimator: data*** $0.9N(0,1) + 0.1N(0,100)$***, escort parameter*** $\sigma^2 = 0.6$

α/n	50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$
0.00	3.173	1.168	1.000	3.178	0.798	1.000	3.254	0.561	1.000	3.266	0.360	1.000
0.01	1.952	0.394	5.734	1.966	0.276	5.327	1.969	0.192	5.529	1.974	0.119	5.470
0.05	1.311	0.160	49.65	1.313	0.107	49.18	1.314	0.074	51.75	1.316	0.048	51.52
0.10	1.176	0.124	130.6	1.176	0.083	141.9	1.177	0.057	156.5	1.177	0.037	160.5
0.20	1.103	0.115	255.8	1.100	0.078	333.4	1.099	0.054	425.8	1.099	0.035	479.2
0.50	1.079	0.140	235.0	1.069	0.096	387.7	1.061	0.067	657.1	1.061	0.042	951.6

As in the location case, we tried to escort the subdivergence estimator with the MLE $\sigma = \mathbf{S}_n$. However, we received only a perfect match with maximum likelihood estimator, showing no robustness whatsoever. This motivated us to plug in a simple and robust estimate of scale called median absolute deviation (MAD), which showed up to be a better choice. For uncontaminated data escorting with MAD works as well as escorting by MLE. For contaminated data we observe again that the estimators perform better than MLE, however we see that the performances are not as good as those of the estimators escorted by $\sigma^2 \in (0.5,1\rangle$. Much better results were received for median absolute deviation where we left out the calibration constant 1.483 (cf. Table 3). Moreover, we compared the subdivergence estimators with MAD itself by plugging it in the empirical relative efficiency formula instead of MLE, and we see that for higher values of α the subdivergence estimators perform even better than the well know robust estimator MAD (cf. Table 4). This outstanding behavior in some cases unfortunately vanishes with higher contamination.

5. References

[1] BRONIATOWSKI, M. – KEZIOU, A. 2006. Minimization of ϕ-divergences on sets of signed measures. In: Studia Scientiarum Mathematica Hungarica, 2006, vol. 43, 403--442.

[2] BRONIATOWSKI, M. – VAJDA, I. 2009. Several Applications of Divegence Criteria in Continuous Families. Research Report No. 2257, Institute of Information Theory and Automation, Prague, 2009.

[3] FRÝDLOVÁ, I. 2004. Minimum Kolmogorov distance estimators. Thesis, Czech Technical University, Prague, 2004.

[4] FRÝDLOVÁ, I. 2009. Modified Power Divergence Estimators: Performances in Location Models. Research Report No. 2258, Institute of Information Theory and Automation, Prague, 2009.

[5] LIESE, F. – VAJDA, I. 2006. On divergences and informations in statistics and information theory. In: IEEE Transactions on Information Theory, 2006, vol. 52, No. 10, 4394--4412.

Table 3: Max $\underline{D}_\alpha$ ***-estimator: data*** $0.9N(0,1)+0.1N(0,100)$***, escort parameter*** $\sigma = MAD/1.483$

α/n	50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$
0.00	3.173	1.168	1.000	3.178	0.798	1.000	3.254	0.561	1.000	3.266	0.360	1.000
0.01	1.936	0.510	5.353	1.940	0.353	5.338	1.933	0.250	5.780	1.941	0.155	5.782
0.05	1.311	0.255	37.65	1.302	0.174	44.29	1.295	0.122	52.83	1.298	0.078	55.49
0.10	1.176	0.194	88.76	1.169	0.131	117.2	1.164	0.092	152.6	1.165	0.059	171.0
0.20	1.098	0.157	178.0	1.095	0.107	263.8	1.090	0.076	390.0	1.091	0.048	491.9
0.50	1.058	0.140	265.8	1.058	0.097	423.1	1.055	0.070	679.9	1.059	0.044	985.3

Table 4: Max $\underline{D}_\alpha$ ***-estimator: data*** $0.9N(0,1)+0.1N(0,100)$***, escort parameter*** $\sigma = MAD/1.483$

α/n	50			100			200			500		
	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$	$m(\tilde{\sigma})$	$s(\tilde{\sigma})$	$eref\ (\tilde{\sigma})$
MAD	0.761	0.129	1.000	0.758	0.090	1.000	0.753	0.065	1.000	0.756	0.042	1.000
0.00	3.173	1.168	0.012	3.178	0.798	0.012	3.254	0.561	0.012	3.266	0.360	0.012
0.01	1.936	0.510	0.065	1.940	0.353	0.066	1.933	0.250	0.070	1.941	0.155	0.068
0.05	1.311	0.255	0.455	1.302	0.174	0.549	1.295	0.122	0.639	1.298	0.078	0.649
0.10	1.176	0.194	1.072	1.169	0.131	1.453	1.164	0.092	1.846	1.165	0.059	2.000
0.20	1.098	0.157	2.149	1.095	0.107	3.270	1.090	0.076	4.718	1.091	0.048	5.751
0.50	1.058	0.140	3.209	1.058	0.097	5.244	1.055	0.070	8.225	1.059	0.044	11.52

Authors' addresses:

Iva Frýdlová, Ing.
FNSPE CTU
Trojanova 13, 120 00 Praha 2
frydlova.iva@seznam.cz

Václav Kůs, Ing. Ph.D.
FNSPE CTU
Trojanova 13, 120 00 Praha 2
vaclav.kus@fjfi.cvut.cz

This work has been supported by the grant SGS OHK4-007/10.

Statistical Properties of Generalized Cramer-von Mises Type Estimate
Statistické vlastnosti zobecněných odhadů Cramer-von Mises typu

Jitka Hanousková, Václav Kůs

Abstract: This paper focuses on estimating via minimum distance method. Two new types of minimum distance estimates are defined, namely with generalized Cramer-von Mises distance and with newly defined Kolmogorov-Cramer distance. Kolmogorov-Cramer estimates are proven to be consistent of the order $n^{-1/2}$ in L_1-norm if the distribution is non-contaminated. Numerical simulation is produced for variously contaminated distributions, and resulting graphs are presented and discussed. The similar simulation comparative study was carried out for the generalized Cramer-von Mises estimate.

Key words: Minimum distance estimate, Consistency, Robustness, Cramer-von Mises

Klíčová slova: Odhady s minimální vzdáleností, konsistence, robustnost, Cramer-von Mises

JEL classification: C13

1. Introduction

This paper introduces a generalized Crame-von Mises distance with a parameter α, and so called Kolmogorov-Cramer distance between empirical and arbitrary distribution function with parameters α, *m*. For non-contaminated case, we have proven consistency and order of consistency of the Kolmogorov-Cramer estimate. Consistency and robustness of these newly defined estimates are explored via computer simulation.

2. Preliminaries

We introduce the notation used throughout the paper. Let λ be a $\sigma-$ finite measure on $(\mathsf{R},\mathcal{B})$, where $\mathcal{B}$ is a borel $\sigma-$field on R. Let $\mathcal{F}_\lambda$ be the set of distributions on $(\mathsf{R},\mathcal{B})$ which are absolutely continuous with respect to a measure λ. Denote by $\mathcal{D}_\lambda$ the set in Banach space $L_1(\mathsf{R},\mathrm{d}\lambda)$ containing densities corresponding to the distribution functions in $\mathcal{F}_\lambda$, by $\mathcal{D}$ its arbitrary nonempty subset and by $\mathcal{F}$ a subset of $\mathcal{F}_\lambda$. Throughout the text $\tilde{F}_n$ denotes the distribution function corresponding to the estimator $\tilde{f}_n$ of the density *f*. Further, let $X_1,..., X_n$ be i.i.d. random sample and F_n denotes the empirical distribution function based on $(X_1,..., X_n)$ with $\nu_n(A)$ representing the empirical measure

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j \le x\}}, \qquad \nu_n(A) = \frac{1}{n}\sum_{j=1}^{n} I_{\{X_j \in A\}}, A \subset \mathsf{R},$$

where $I_{\{X_j \le x\}}$ means indicator of the event $\{ X_j \in (-\infty, x]\}$ and $I_{\{X_j \in A\}}$ is the indicator of $\{ X_j \in A \}$.

Definition 1. We say that an estimate $\tilde{f}_n$ of a density $f \in \mathcal{D}$ is *minimum D distance estimate* iff the corresponding distribution function $\tilde{F}_n \in \mathcal{F}$ satisfies the condition:

$$D(\tilde{F}_n, F_n) = \inf_{F \in \mathrm{F}} D(F, F_n) \qquad \text{a.s.}$$

Every metric distance D on $\mathcal{F}(\mathsf{R})$ defines pseudometric ρ_D on $\mathcal{D}_\lambda$ in this sense: $\rho_D(f,g) = D(F,G)$, where $F,G \in \mathcal{F}_\lambda$ are the distribution functions corresponding to densities $f,g \in \mathcal{D}_\lambda$.

Definition 2. We say that an estimate $\tilde{f}_n$ of a density $f \in \mathcal{D}$ is consistent in given ρ_D distance, (in expected ρ_D distance), iff $\rho_D(\tilde{f}_n, f) \to 0$ a.s., (iff $E\rho_D(\tilde{f}_n, f) \to 0$). We say that estimate $\tilde{f}_n$ is consistent of the order $r_n \to 0$ in the distance ρ_D, respective in expected ρ_D distance, iff $\rho_D(\tilde{f}_n, f) = O_p(r_n)$, respective iff $E\rho_D(\tilde{f}_n, f) = O(r_n)$.

We define generalized Cramer-von Mises distance with parameter α ($\rho_{CM,\alpha}$) as:

$$\rho_{C-M,\alpha}(f,g) = \int_{\mathsf{R}} |F(x) - G(x)|^{\alpha} \, \mathrm{d}G(x).$$

Further, we define a new type of minimum distance estimator. We define so called Kolmogorov-Cramer distance with parameters α, m ($\rho_{KC\alpha,m}$) between empirical distribution function and arbitrary distribution function in the following way.

Definition 3. Let $(x_1,\ldots, x_n)$ be a realization of random vector $(X_1,.., X_n)$, α a nonnegative real parameter, and F arbitrary distribution function. Then we define the sequence $(G_i)_1^{2n}$ by

$$G_i = |F_n(x_i) - F(x_i)|^{\alpha} \quad \text{for } i = 1,\ldots,n,$$
$$G_{2n+1-i} = |F_{n-}(x_i) - F_{-}(x_i)|^{\alpha} \quad \text{for } i = 1,\ldots,n,$$

where $F_{n-}(x_i)$ and $F_{-}(x_i)$ are left hand side limits for $x \to x_i$. Then Kolmogorov-Cramer distance is defined as

$$\rho_{CM\alpha,m}(F_n, F) = \frac{1}{m} \sum_{i=1}^{m} G_{(i)},$$

where $G_{(i)}$ denotes descending ordering of G_i, $i = 1,\ldots,2n$. In other words, the Kolmogorov-Cramer distance is one over m times sum of m largest of G_i, $i = 1,\ldots,2n$. We call the distance Kolmogorov-Cramer, because for m=1 it converts to Kolmogorov distance to the power α and the expression is inspired by (generalized) Cramer-von Mises distance for the case G=F_n.

3. Consistency in L_1-norm

In [1], the conditions for the consistency and consistency of the order $n^{-1/2}$ in the L_1-norm within the Kolmogorov estimates are presented. Conditions are based on domination relations between Kolmogorov distance and total variation distance. Furthermore, sufficient condition for the domination relation is proven ibid. We use this results for exploring the consistency and the order of consistency of our minimum distance estimates where the distances are differing from the Kolmogorov distance. If for a given distance D it holds both inequalities $D \le h_1(d_K), d_K \le h_2(D)$, where h_1, h_2 are two real functions continuous in zero point with zero value in zero argument and if there exist positive constants K_i such that $h_i(x) \le K_i x$ in a neighborhood of zero and moreover the domination relation stated in [1] is fulfilled for family $\mathcal{D} \subset \mathcal{D}_\lambda$, then the minimum D distance estimates of densities from $\mathcal{D}$ are consistent of the order $n^{-1/2}$ in L_1-norm and the expected L_1-norm. The upper mentioned inequalities are fulfilled e.g. for Discrepancy and Lévy distance. (See [6] in details, for specific inequalities derived in spaces of probability densities, see [3].) In case that the functions h_i could not be constrained by $K_i x$, it is possible to prove the consistency in L_1-

norm, but the order of consistency $n^{-1/2}$ need not be preserved. Thanks to proper inequalities between the Kolmogorov distance and Kolmogorov-Cramer distance we are able to prove consistency and consistency of the order $n^{-1/2}$ for the Kolmogorov-Cramer estimate. It holds that

$$\rho_{CM\alpha,m} \le (\rho_K)^{\alpha} \quad \text{and} \quad \rho_{CM\alpha,m} \ge 1/m(\rho_K)^{\alpha} ,$$

where $\rho_K(f,g) = \sup\{|F(x)-G(x)|, x \in R\}$ stands for the Kolmogorov distance. Further, we are able to prove consistency of this estimate in the case that *m* depends on sample size *n*, too. In detail, if $m = n^{\beta}$, where $\beta \le \alpha/2$, then the estimate is consistent of the order n^{γ}, where $\gamma = 1/2 + \beta/\alpha$. For generalized Cramer-von Mises distances the desired inequalities leading to optimal $n^{-1/2}$ consistency have not been proved. Nevertheless, we hope for consistency of these estimates, so we produce numerical simulation to ascertain. Via simulation we study robustness of both mentioned estimates, as well.

4. Numerical simulation

We consider normal distribution N(0,1) contaminated by normal distribution N(0,100) and explore consistency and robustness. For contaminated case, we have no theoretical results guaranteeing us the consistency, thus this case is explored only via simulation. Results for normal distribution N(0,1) with 5%, 10%, 15%, and 35% contamination by normal distribution N(0,100) along with non-contaminated case are presented.

Figure 1 presents consistency of generalized Cramer- von Mises estimate with various values of α (α=0.5, 1, 1.5, 2) for the non-contaminated distribution and 5%, 10%, and 15% contaminations. For other choices of α the graphs are of the same shape, and they are not presented here. As can be seen from the graphs the heavier contaminated sample is the worse the consistency is getting. Nevertheless generalized Cramer-von Mises estimate preserves relatively good consistency even under contamination. Thus, this type of estimate shows robust behaviour.

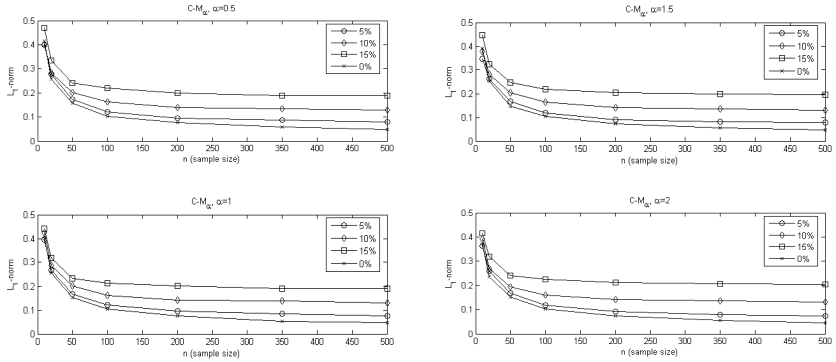


Figure 1: L_1-error of greneralized Cramer-von Mises estimate.

Regarding to the robustness, we seek for the choice of parameter α which produces the best estimate, in the sense of robustness and consistency as well. Figure 2 presents average absolute error of estimated parameter (σ) with respect to the true value of the parameter (σ_0) for non-contaminated distribution, 15%, and 35% contamination.

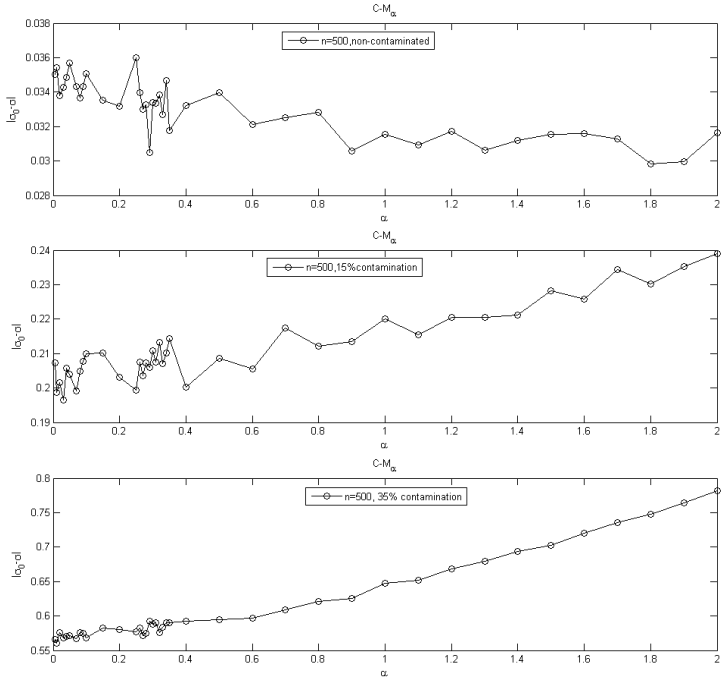


Figure 2: Average absolute error of estimated parameter with respect to the true value of the parameter $|\sigma_0-\sigma|$ of various types of generalized Cramer-von Mises estimate.

As can be seen from the Figure 2, the smaller parameter α we chose the more robust estimator we obtain if the distribution is contaminated. And for non-contaminated case this choice does not cause a collapse (in efficiency). The average absolute error is a little bit greater then for the choices near α=2 (original Cramer-von Mises estimate), but still in tolerable range.

Figure 3 shows consistency of Kolmogorov-Cramer estimate with parameters α=1.7 and *m*=20. For the Kolmogorov-Cramer estimate we have proven consistency and $n^{-1/2}$ order of consistency for non-contaminated distribution. As the graph shows, consistency is preserved even under 5% contamination but under heavier contamination this estimate does not posses consistency anymore. Graphs for other choices of parameters α, *m* are not presented here since they are of almost the same shape.

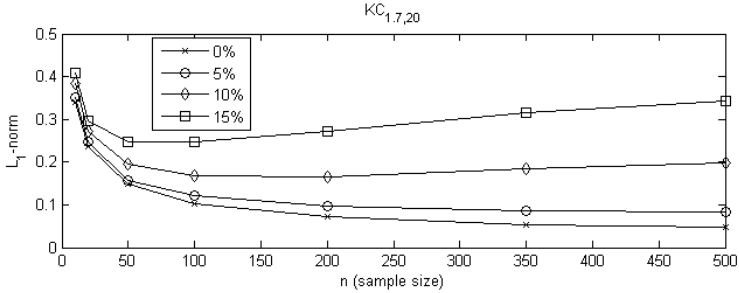


Figure 3: L_1-error of Kolmogorov-Cramer estimate with parameters α=1.7 and m=20.

Figure 4 presents comparison of Kolmogorov, Kolmogorov-Cramer (*m*=20, 50, 90), and generalized Crame-von Mises estimate for various choices of parameter α. The best of all is the generalized Cramer-von Mises estimate for all examined choices of α and under any contamination. Due to the resemblance of the graphs, only non-contaminated and 15% contamination cases are presented here.

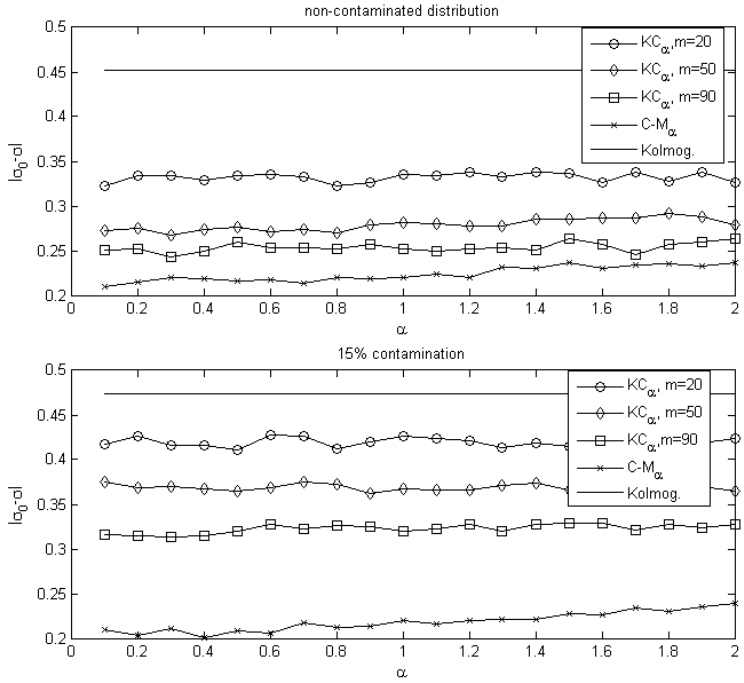


Figure 4: Comparison of absolute average error of estimated parameter with respect to the true parameter $|\sigma_0-\sigma|$ of all examined estimates for various choices of parameter α.

5. Conclusion

By means of numerical simulation we explored consistency of generalized Cramer-von Mises estimate for non-contaminated and contaminated Normal distribution. For the Kolmogorov-Cramer estimate we have theoretical result guaranteeing the consistency and order of consistency even if the parameter m depends on sample size n. For contaminated distribution we produced simulation which shows that this estimate preserves consistency under weak contamination. Further, robustness of generalized Cramer-von Mises, Kolmogorov-Cramer, and Kolmogorov estimates was compared.

6. References

[1] KŮS V. 2004. Nonparametric density estimates consistent of the order of $n^{-1/2}$ in the L_1-norm. In: Metrika, 2004, s. 1-14.

[2] DEVROYE, L. - GYÖRFI, L. - LUGOSI, G. 1996: A Probabilistic Theory of Pattern Recognition, New York: SPRINGER, 1996.

[3] GIBBS, A. L., SU, F. E., 2002. On choosing and bounding probability metrics. In: International Statistical Review,70, 2002, s.419-435.

[4] FRÝDLOVÁ, I. 2004. Odhady pravděpodobnostních hustot s minimální Kolmogorovskou vzdáleností, Diplomová práce, FJFI ČVUT Praha, 2004.

[5] DEVROYE, L. - GYÖRFI, L 1985. Nonparametric density Estimate, the L_1 – view, New York: Wiley, 1985.

[6] HANOUSKOVÁ, J.: Asymptotické vlastnosti odhadů s minimální vzdáleností, Diplomová práce, FJFI ČVUT Praha, 2009.

Adresa autora (-ov):

Jitka Hanousková, Ing
Katedra matematiky, FJFI ČVUT
Trojanova 13, 120 00 Praha 2
hanoujit@fjfi.cvut.cz

Václav Kůs, Ing, PhD.
Katedra matematiky, FJFI ČVUT
Trojanova 13, 120 00 Praha 2
vaclav.kus@fjfi.cvut.cz

This work has been supported by the grant SGS OHK4-007/10.

New Classification Attributes for Signal Separation in Acoustic Emission Processing

Nové klasifikační atributy pro separaci signálů v akustické emisi

Václav Kůs, Zuzana Farová, Jan Tláskal

Abstract: We deal with the classification of acoustic emission random signals by means of fuzzy clustering, model-based statistical clustering and support vector machines. The signals are compared by means of newly introduced attributes obtained directly from the signals and from normed frequency spectra such as ϕ-divergence distance measure originated from information theory and statistics. We mention the well-known divergences (Kullback, Hellinger, χ^2, Power) and we introduce several of its new modifications and extensions as generalized Hellinger divergences. These new families of divergences open new research possibilities in the area of statistical treatment of random signals. We realize experiments to test the proposed classification attributes and also consider industrial data from the real life.

Key words: Classification attributes, ϕ-divergences, Signal separation, Data processing.

Klíčová slova: Klasifikační atributy, ϕ-divergence, separace signálů, zpracování dat.

JEL classification: C61

1. Newly designed attributes for signal separation

All processing laboratory measured acoustic signals were detected through the piezo-ceramic sensors attached to the thin metal plate or have been provided by DAKEL from various places of their measurements. Emitted signals were measured and stored on a computer by means of measuring device DAKEL-XEDO 5 in 12-bit accuracy and 4 MHz sampling rate. The resolution 12-bit means that the measuring sampling apparatus was able to distinguish the voltage in the interval [-2048mV, 2048mV]. We generated several versions of acoustic sources in the central part of experimental plate.

We perform our cluster analysis with certain attributes computed also from the normalized signal spectrums X_f found through discrete Fourier transform. The spectrum is normalized to one in order to obtain the estimate of spectral density of the signal $\{X_t\}_{t=0}^{T-1}$, where T stands for digital length of the emission event (in our case $T = 6144$). The normed spectrum is given by $\tilde{S}(f) = |X_f| / \sum_{f=0}^{T-1} |X_f|$. Two examples of acoustic emission signals with corresponding spectrum densities can be seen in Figure 1.

We compute several attributes from the normed spectrums or directly from the signals and then we use these parameters in the methods of classification specified further. We designed the following **classification attributes** W_α, Q_β, Z_c, D_ϕ:

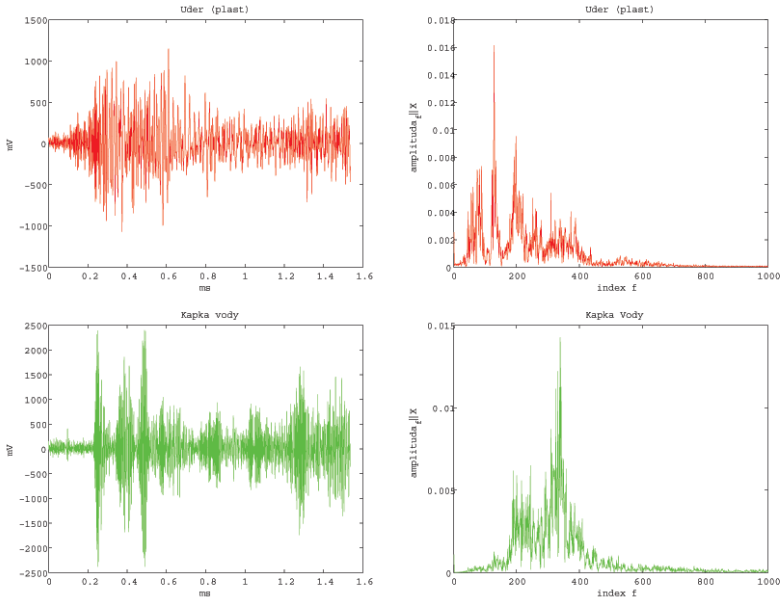


Figure 1: Examples of acoustic emission signals and spectrums.

$$W_\alpha = \arg\min_{l\in[0,T-1]} \sum_{f=0}^{T-1} |\, l-f \,|\, |X_f| - |\overline{X}_f| \,|^\alpha, \quad \alpha\in[1,\infty),$$

$$Q_\beta = \min\{F\in[0,T-1]: \sum_{f=0}^{F} |X_f| \geq \beta\}, \quad \beta\in(0,1),$$

$$Z_c = \sum_{t=\overline{1}}^{T-1} \delta(x_t), \quad \delta(x_t) = \begin{cases} 1 & \text{if} sgn(x_t x_{t+1}) = -1 \\ 0 & \text{otherwise} \end{cases}, \quad c\in(0,1),$$

$$D_\phi = \sum_{f=0}^{T-1} \tilde{S}(f)\phi\left(\tilde{S}^{refer}(f)/\tilde{S}(f)\right), \quad \phi:(0,\infty)\to \mathsf{R}, \phi(1)=0,$$

where $|\overline{X}_f|$ denotes the arithmetic mean $|\overline{X}_f| = \sum_{f=0}^{T-1} |X_f|/T$,

$$J_\gamma = \{j\in[0,T-1]: |X_j| \geq \gamma max_{f\in[0,T-1]} |X_f|\}, \quad \gamma\in(0,1),$$

and $\tilde{S}^{refer}$ is a normed reference spectrum, $\tilde{S}^{refer}(f) = \sum_{i=1}^{m} |\tilde{S}^i(f)|/m$. Here m denotes the number of observations from one sort of acoustic emission signals, $\tilde{S}^i(f)$ are individual realizations of the normed spectrum $\tilde{S}(f)$.

In this paper we work with the specific classification attributes: W_2 (denoted as W), $Q_{0.33}$, $Z_{\frac{1}{20}}$ (denoted as Z_c), D_ϕ discrete form of the generalized Hellinger divergence with blending parameter $\beta = 0.5$ (denoted as *HDiv*, see Section 3).

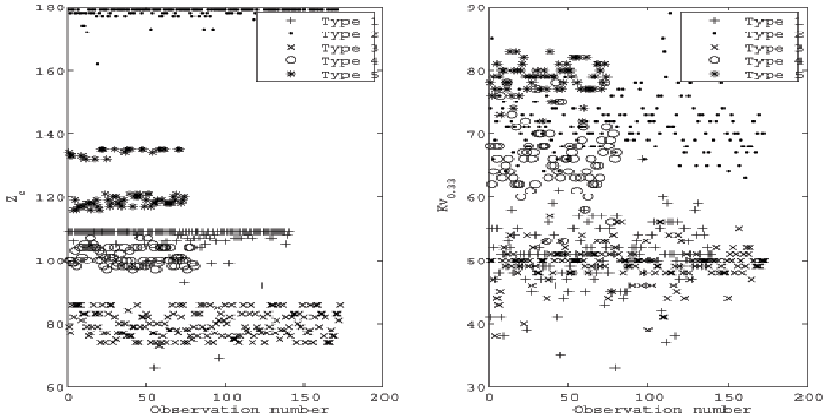


Figure 2: Signal separation by means of attributes Z_c and $Q_{0.33}$.

We measured five different types of acoustic emission signals in a laboratory environment. As an example, some of the above mentioned spectral attributes were computed for each signal and the separation, by means of the proposed attributes, can be seen in Figure 2. Here, in the case of our laboratory data set, the newly introduced parameter Z_c separates the signal clusters very well, five distinct sets of Z_c values are apparently present. The other signal attributes separate the clusters in a similar way, i.e. some signal types remained overlapped.

2. Novel signal attributes based on ϕ-divergences

Let (X, A) be a measurable space and let P be the set of all probability spectral measures on (X, A). If $P \in P$ is dominated by a σ--finite measure μ on (X, A), $p = \mathrm{d}P/\mathrm{d}\mu$ denotes the spectral density of the acoustic signal (so-called Radon-Nikodym derivative of P with respect to μ). Now let $P, Q \in P$, $\{P, Q\} = \mu$, $p = \mathrm{d}P/\mathrm{d}\mu$ and $q = \mathrm{d}Q/\mathrm{d}\mu$ are two signal spectral densities. *ϕ--Divergence* of P and Q is a function $D_\phi : P \times P \to [0, \infty]$ defined by

$$D_\phi(P,Q) = \int_X q\phi\left(\frac{p}{q}\right)\mathrm{d}\mu,$$

where $\phi : (0, \infty) \to \mathsf{R}$, $\phi(1) = 0$, $\phi(t)$ is convex on $(0, \infty)$ and strictly convex at $t = 1$. We put $q\phi(p/q) = q\phi(0)$ if $p = 0$ and $q\phi(p/q) = p\phi(\infty)/\infty$ if $q = 0$, where $\phi(0) := \lim_{t \to 0_+} \phi(t)$ and $\phi(\infty)/\infty := \lim_{t \to \infty} \phi(t)/t$ with the convention „$0 \cdot \infty = 0$“. It is known that D_ϕ are all reflexive (i.e. $D_\phi(P, P) = 0$), $\phi(0) + \phi(\infty)/\infty > 0$ and range is $0 \le D_\phi(P, Q) \le \phi(0) + \phi(\infty)/\infty, P, Q \in P$, where the upper bound is achieved if P, Q are two singular spectral measures. Further, $D_\phi(P, Q)$ are all invariant with respect to the linear transform of the form $\tilde{\phi}(t) = \phi(t) - \phi'_+(1)(t-1)$, $t \in (0, \infty)$, where $\phi'_+(1)$ denotes derivatives of ϕ at $t = 1$ from the

right. Thus every ϕ-function has its nonnegative version $\tilde{\phi}$ with the same ϕ-divergence and $\tilde{\phi}_+'(1)=0$. For advanced theory of ϕ-divergences see [1] or [2].

Now, let $\psi:(0,\infty)\to\mathsf{R}$ be a convex or concave function on $(0,\infty)$, with the property of being twice differentiable at $t=1$ with $\psi''(1)\neq 0$. Then

$$\bar{\phi}(t)=(\psi(t)-\psi(1)-\psi'(1)(t-1))/\psi''(1),\qquad t\in(0,\infty), \tag{1}$$

is a fully normalized divergence function, i.e. $\bar{\phi}(t)$ is nonnegative, twice differentiable at $t=1$, $\bar{\phi}'(1)=0$, and $\phi''(1)=1$. Table 2 illustrates how some well-known divergences arise from very simple ψ-functions via construction (1).

Table 1: ψ-functions for standard ϕ-divergences

	Kullback (I)	**Pearson**(χ^2)	**Neyman** (χ^2)	**Le Cam** (LC^2)
$\psi(t)$	$\ln t$	t^2	$\frac{1}{t}$	$\frac{1}{1+t}$
$\bar{\phi}(t)$	$-\ln t+t-1$	$(t-1)^2/2$	$\frac{(t-1)^2}{2t}$	$\frac{(t-1)^2}{t+1}$
	Hellinger (H^2)	**Power** (I_a)	**Tuned** (D_a)	
$\psi(t)$	$\sqrt{t}$	$t^a, a\neq 0,1$	$t^{(a+1)/2}, a\neq\pm 1$	
$\bar{\phi}(t)$	$2(\sqrt{t}-1)^2$	$\frac{t^a-a(t-1)-1}{a(a-1)}$	$4\frac{t^{\frac{a+1}{2}}-(a+1)(t-1)/2-1}{(a+1)(a-1)}$	

3. Generalized Hellinger divergence

The presented method of construction (1) can be combined with standard transformation techniques preserving convexity of transformed functions. Let start, for all $b\geq 0$, with the ψ-function $\psi_{a,b}(t)=(1-t^{2a})/(b+t^a)$, $b\geq 0$, $a\neq 0,1$. We employ the convex transformation $f(y)=y^2$ to obtain $\psi^2_{a,b}(t)$. Since $\psi_{a,b}$ is not nonnegative on $(0,\infty)$ then the transformation $f(y)=y^2$ need not preserve convexity nor concavity and it have to be treated separately. Let us consider only the special case for $a=1/2$. Then the function

$$\phi_b(t):=\psi^2_{\frac{1}{2},b}(t)=\left(\frac{1-t}{b+\sqrt{t}}\right)^2,\qquad t\in(0,\infty),$$

is convex for all $b\geq 0$ and construction (1) provides us with the fully normalized divergence function $\bar{\phi}_b(t)=(b+1)^2\phi_b(t)/2$. The blending parameter $\beta=1/(1+b)\in[0,1]$ leads to

$$\bar{\phi}_\beta(t)=\frac{1}{2}\left(\frac{1-t}{1-\beta+\beta\sqrt{t}}\right)^2,\qquad t\in(0,\infty),\quad \beta\in[0,1],$$

with the corresponding *generalized Hellinger divergence*

$$H_\beta(P,Q)=\frac{1}{2}\int\frac{(p-q)^2}{(\beta\sqrt{p}+(1-\beta)\sqrt{q})^2}\mathrm{d}\mu,\qquad P,Q\in P.$$

The particular cases are Pearson's $\chi^2(P,Q)/2$ for $\beta = 0$ $(b \to \infty)$, Neyman's $\chi^2(Q,P)/2$ for $\beta = 1$ $(b = 0)$, and Hellinger's $2H^2(P,Q)$ for $\beta = 1/2$ $(b = 1)$. The function $\overline{\phi}_\beta(t)$ is strictly convex on $(0,\infty)$ for all $\beta \in [0,1]$ and all the blended divergences $H_\beta(P,Q)$ for $\beta \in (0,1)$ are bounded. For $\beta \in [0,1]$ the skew symmetry about $\beta = 1/2$ takes place, $H_\beta(P,Q) = H_{1-\beta}(Q,P)$, $P,Q \in P$, which provides the only symmetric divergence $H_{\frac{1}{2}}(P,Q) = 2H^2(P,Q)$ in this class.

4. Signal separation for the real data sets

We deal with the classification of acoustic emission signals by means of Fuzzy Clustering (FC), Model-Based Clustering (MBC) and Support Vector Machines (SVM). These methods belong to the different groups of classification techniques.

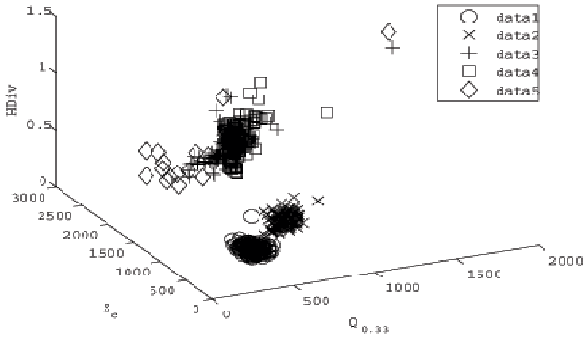


Figure 3: Parameters $Q_{0.33}$, Z_c and $HDiv$ for the real data sets.

The FC method is based on optimization of so-called objective function Q in the form of weighted sum of quadratic distances between individual signal attributes x_i and a set of the so-called prototypes (i.e. the representative samples) $v_1, v_2, \ldots, v_c$ of c clusters using the fuzzyfication factor and partition matrix $U = [u_{ik}]$, which allocates the signals into the k-th cluster. We employ numerical iterative fuzzy algorithm to find the minimum of objective function Q under the restrictions. This algorithm is based on stepwise calculation of the prototypes of the clusters and the partition matrix until certain convergence criterion is satisfied. More details can be found in [3].

The MBC method supposes that the data belong to a finite mixture of normal densities, where each component represents one cluster. It means that we deal with the finite convex combination of probability densities given the sample of signal attributes $\mathbf{x} = (x_1, \ldots, x_n)$, $x_i \in \mathsf{R}^d$. The best normal mixture fit is performed through the maximum likelihood principle and it is based on a general iterative EM algorithm for finding maximum of the likelihood function under consideration. It manages to work with uncomplete data sets in the form $y_i = (x_i, z_i)$, which contain the so-called unobserved data z_i of the binary values

referring to the seeking membership of the corresponding signal to a certain cluster. See [4] for more detailed analysis and proofs.

The SVM method requires learning computer based process, it means we need some training data set adjusting the procedure. Let us consider that we are endowed by a training data set $(x_i, y_i) \in \mathsf{X} \times \{+1, -1\}$, where x_i are again the observed signal attributes computed from signal or spectral densities, and y_i are separating labels. The classification into two clusters is represented by two values of labels, $+1$ or -1. The method separates data by searching for optimal separating hyperplanes between clusters requiring the margin along the hyperplane to be as wide as possible. The optimal separating hyperplane is designed by means of binary decision tree with iterative minimization algorithm under constraints in each node of the decision tree. See [5] for the detailed SVM method procedure.

We apply the three methods FC, MBC, and SVM with the proposed classification attributes on the real data set, which contains 5 types of acoustic emission sources originated from the measurements handled by DAKEL through the device Xedo-5 at the different places of interest, i.e.

Table 2: Percentage separation successes for the real data set.

Method	Specification of method	$Q_{0,33}, Z_c$	$Q_{0,33}, Z_c, HDiv$
FC	Classical non-penalized	76,5%	76,5%
MBC	5 clusters	88,2%	**89%**
SVM	C=5, linear	51,4%	**83,7%**
SVM	C=5, polynomial degree 2	85,6%	83,9%
SVM	C=5, erbf degree 1	**90,8%**	**90,8%**

- Data1: Measurement of the CAVITATION
- Data2: Experiment with the PLASMATRON
- Data3: Process of WELDING
- Data4: Pressure test on BOILER VESSEL
- Data5: Rigidity test for LAMINATE PLATE

Classification successfulness for the real data set is listed in Table 2 for the combinations of parameters $(Q_{0.33}, Z_c)$ and $(Q_{0.33}, Z_c, HDiv)$. The combination of parameters $(Q_{0.33}, Z_c, HDiv)$ for the real data set is shown in Figure 3. We achieve the highest percentage success by using the method SVM, **90,8%**, however, success by using the method MBC is also remarkable, **89%** for parameters $Q_{0,33}, Z_c, HDiv$. Moreover, it can be seen from Table 2 that the success rate is much higher when using linear SVM with parameter $HDiv$ added.

5. References

[1] VAJDA, I. 1989. Theory of Statistical Inference and Information. Kluwer, Boston.

[2] KŮS, V.; MORALES, D.; VAJDA, I. 2008. Extension of the Parametric Families of Divergences Used in Statistical Inference. *Kybernetika*, 44/1, 95 - 112, 2008.

[3] PEDRYCZ, W. 2005. Knoledge-Based Clustering, From Data to Information Granules. Wiley-Interscience, New Jersey, 2005.

[4] FRALEY, C.; RAFTERY, A. E. 2002. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Jourtnal of the American Statistical Association*, 97, 611-631, 2002.

[5] SCHÖLKOPF, B.; SMOLA, A. J. 2002. Learning with Kernels -- Support Vector Machines, Regularization, Oprimalization and Beyond. The MIT Press, 2002.

Adresa autora (-ov):

Václav Kůs, Ing, PhD.
Katedra matematiky, FJFI ČVUT
Trojanova 13
120 00 Praha 2
vaclav.kus@fjfi.cvut.cz

Zuzana Farová, Ing.
KM FJFI ČVUT
Trojanova 13
120 00 Praha 2
zuzana.farova@klikni.cz

Jan Tláskal, Ing.
KM FJFI ČVUT
Trojanova 13
120 00 Praha 2
tlaskalik@email.cz

This work has been supported by the grant SGS OHK4-007/10.

Analýza bloku Právo duševného vlastníctva vo výkaze INOV 1-99 podsystémom PivotTable systému EXCEL za rok 2006

Analyse block Intellectual property rights at report INOV 1-99 subsystem PivotTable system EXCEL within a year 2006

Jozef Chajdiak

Abstract: In contributions myself to present technique finding frequency about intellectual property rights through the medium PivotTable in Excel. Finding are frequency on basis counts statistical unit and on basis sustain turnover behind year 2006. Most significant digit frequency reaches registration safety marks.

Abstrakt: V príspevku sa prezentuje postup zisťovania frekvencií pri právach duševného vlastníctva pomocou PivotTable v Exceli. Zistené sú frekvencie na báze počtu štatistických jednotiek a na báze úhrnu obratu za rok 2006. Najvyššie frekvencie dosiahla registrácia ochrannej značky.

Key words: Intellectual property, patent, industrial design, trademark, copyright, distribution, frequency.

Kľúčové slová: duševné vlastníctvo, patent, priemyselný vzor, ochranná známka, autorské právo, PivotTable, triedenie, početnosť.

JEL classification: C13, O34

1. Úvod

Právo duševného vlastníctva predstavuje dôležitú súčasť procesov inovácií. Obsahuje žiadosti o patent, registráciu priemyselných vzorov, registráciu ochrannej známky a uplatnenie autorského práva.

2. Východiskové údaje

Vlastné údaje vykázané vo výkaze Štatistické zisťovanie o inováciách za rok 2006 Inov 1-99 sme zredukovali v počte premenných len na túto analýzu potrebné údaje (hárok VS2010). Prvé riadky upraveného súboru sú nasledujúce:

	A	B	C	D	E	F	G
1	ID	Nace_pro	TURN06m	PROPAT	PRODSG	PROTM	PROCP
2	1	73_74	658913	0	0	0	0
3	2	DI	448285	0	0	0	0
4	3	DH	35617307	0	1	1	0
5	4	DH	310123073	1	1	1	1
6	5	DH	34967052	0	0	0	0
7	6	20_21	71002014	0	0	1	0

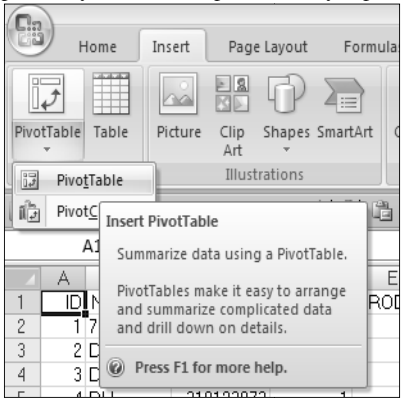
Súbor obsahuje premenné:
ID – poradové číslo pozorovania,
Nace_pro – kód NACE,
TURN06m - obrat (v mil. Sk),
PROPAT – žiadosť o patent (0 – nie, 1 – áno),
PRODSG – registrácia priemyselného vzoru (0 – nie, 1 – áno),
PROTM – registrácia ochrannej známky (0 – nie, 1 – áno),
PROCP – uplatnenie autorského práva (0 – nie, 1 – áno).

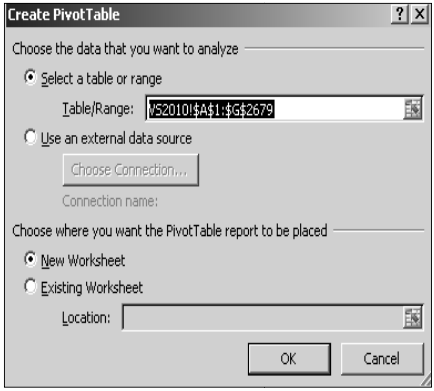
3. Úlohy analýzy

Úlohou je údaje vytriediť a určiť absolútne a relatívne triedne početnosti na báze počtu štatistických jednotiek a na báze obratu štatistických jednotiek.

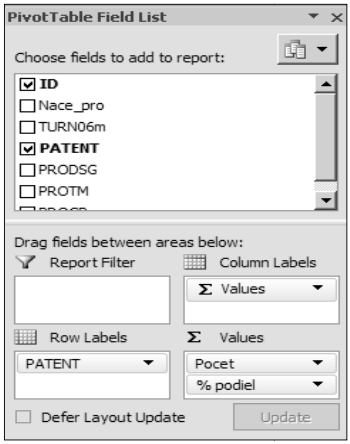
4. Riešenie úloh analýzy pomocou PivotTable v Exceli

Podsystém PivotTable (Kontingenčná tabuľka) aktivujeme z hárku VS2010 postupnosťou príkazov: Insert/PivotTable/PivotTable/ (ľavá časť výstupu) a potvrdením ponúkaných možností (pravá časť výstupu):





V časti PivotTable Field List presunieme premennú ID z políčka Choose fields to add to report do políčka Values, zmeníme z verzie SUM na COUNT (počet) a premenujeme na Počet – zabezpečí výpočet absolútnych triednych početností. Premennú ID do políčka Values presunieme ešte raz, z verzie SUM zmeníme na COUNT, verziu Normál zmeníme na % of col a premenujeme na %podiel – zabezpečí výpočet percentuálnych triednych početností. Do políčka Row Labels presunieme premennú PROPAT, ktorú sme premenovali na PATENT. Výsledky triedenia sú na výstupe vpravo.



	Data	
PATENT	Pocet	% podiel
0	2627	98,1%
1	51	1,9%
Grand Total	2678	100,0%

Postupne z poľa Row Labels vysunieme príslušnú premennú a vložíme novú premennú, ktorú premenujeme na VZOR, ZNAMKA a PRAVO. Výsledky triedenia sú nasledujúce:

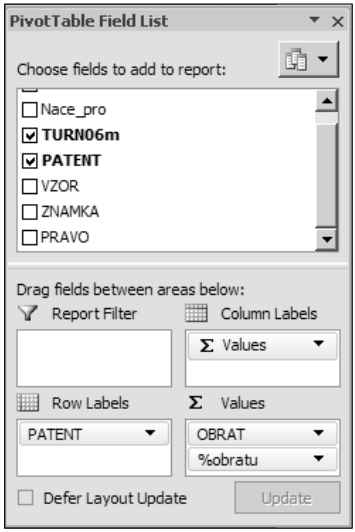
VZOR	Pocet	% podiel
0	2604	97,2%
1	74	2,8%
Grand Total	2678	100,0%

ZNAMKA	Pocet	% podiel
0	2414	90,1%
1	264	9,9%
Grand Total	2678	100,0%

PRAVO	Pocet	% podiel
0	2599	97,1%
1	79	2,9%
Grand Total	2678	100,0%

Vidíme, že najčastejšou aktivitou je skoro 10% podiel zaregistrovaní ochrannej známky (premenná ZNAMKA). K 3% sa blíži zaregistrovanie priemyselného vzoru (VZOR; 2,8%) a uplatnenie autorského práva (PRAVO; 2,9%). Najnižšia frekvencia je pri žiadostiach o patent (PATENT; 1,9%).

Iný pohľad na problematiku dáva určenie frekvencií vychádzajúce z dosiahnutého obratu. Určenie frekvencií na báze obratu je podobné ako určenie frekvencií na báze počtu firiem. V políčku Values vymeníme Počet za OBRAT a %podiel za %obratu.



PATENT	OBRAT	%obratu
0	55 284 819 990	90,11%
1	6 071 064 518	9,89%
Grand Total	61 355 884 508	100,00%

VZOR	OBRAT	%obratu
0	55 993 540 731	91,26%
1	5 362 343 777	8,74%
Grand Total	61 355 884 508	100,00%

ZNAMKA	OBRAT	%obratu
0	41 657 824 761	67,90%
1	19 698 059 747	32,10%
Grand Total	61 355 884 508	100,00%

PRAVO	OBRAT	%obratu
0	49 910 536 589	81,35%
1	11 445 347 919	18,65%
Grand Total	61 355 884 508	100,00%

Vidíme, že Práva duševného vlastníctva sa uplatňujú vo výraznej vyššej frekvencii založenej na báze úhrnného obratu (od 8,74% po 32,10%) ako na báze počtu vykazujúcich jednotiek (od 1,9% po 9,9%). Zaregistrovanie ochrannej známky je u 32,1% obratu firiem a naopak ochrana priemyselného vzoru je u 8,74% obratu firiem.

4. Záver

Najvyužívanejšou formou ochrany práv duševného vlastníctva je registrácia ochrannej známky (9,9% firiem, resp. 32,1% obratu). Ďalej zo zistených frekvencií môžeme konštatovať, že jednotky s vyšším obratom majú vyššiu úroveň ochrany práv duševného vlastníctva.

5. Literatúra

[22] CHAJDIAK, J. 2009. ŠTATISTIKA V EXCELI 2007. Bratislava: Statis, 2009, 304 s. ISBN 978-80-85659-49-8

[23] CHAJDIAK, J. 2010: Štatistika jednoducho (3. vydanie). Bratislava: Statis 2010, 194 strán. ISBN 978-80-85659-60-3

[24] www.statistics.sk

[25] Výkaz Inov 1-99

[26] Súbor anonymizovaných údajov CIS2006SK. Prameň ŠÚ SR

Adresa autora:

Jozef Chajdiak, Doc., Ing., CSc.
Ústav manažmentu STU
Bratislava
Jozef.Chajdiak@stuba.sk

Vypracované v rámci riešenia úlohy VEGA č.1/0536/10 „Inovácia ako strategický základ zvyšovania konkurenčnej schopnosti SR“

Materiály Valného zhromaždenia členov SŠDS

Uznesenia Valného zhromaždenia členov SŠDS, konaného 7. októbra 2010.

Návrhová komisia pracovala v zložení: Vaňo B., Stankovičová I., Podmanická Z.

Volebná komisia pracovala v zložení: Úradníček V., Mach P., Bleha B.

Návrhová komisia pripravila návrh uznesení, ktoré boli Valným zhromaždením schválené.

1. Valné zhromaždenia členov SŠDS schvaľuje:

a) Správu o činnosti Správa o činnosti Slovenskej štatistickej a demografickej spoločnosti za obdobie rokov 2006 – 2010
b) Správu revíznej komisie za obdobie rokov 2006 - 2010
c) Hlavné úlohy Slovenskej štatistickej a demografickej spoločnosti na obdobie rokov 2010 – 2014

2. Valné zhromaždenie členov SŠDS zvolilo tajným hlasovaním členov Výboru SŠDS na obdobie rokov 2010 – 2014,

menovite:

Bernadič František, Bleha Branislav, Cuper Ján, Chajdiak Jozef, Ivančíková Ľudmila, Janusová Anna, Katina Stanislav, Komorník Jozef, Korony Samuel, Kusendová Dagmar, Luha Ján, Mach Peter, Mládek Jozef, Nánásiová Oľga, Páleník Viliam, Pastor Karol, Potocký Rastislav, Potančoková Michaela, Radvanský Marek, Stankovičová Iveta, Stehlíková Beáta, Šprocha Branislav, Tkáč Michal, Tirpáková Anna, Úradníček Vladimír, Vojtková Mária, Wimmer Gejza .

3. Valné zhromaždenie členov SŠDS zvolilo tajným hlasovaním členov Revíznej komisie na obdobie rokov 2010 – 2014,

menovite:

Cár Mikuláš, Kanderová Mária, Šoltés Erik.

4. Valné zhromaždenie členov SŠDS schvaľuje platnosť Stanov SŠDS. Nové znenie bude publikované v odbornom časopise SŠDS FORUM STATISTICUM SLOVACUM 5/2010.

Dňa 7. 10. 2010 v Únovciach nad Váhom, okr. Galanta

Boris Vaňo
Iveta Stankovičová
Zuzana Podmanická

Slovenská štatistická a demografická spoločnosť
Miletičova 3, 824 67 Bratislava

Zloženie Výboru a Revíznej komisie SŠDS na obdobie 2010-2014

Predseda:
Doc. Ing. Jozef Chajdiak, CSc., ÚM STÚ Bratislava

Vedecký tajomník:
RNDr. Ján Luha, CSc., LF UK Bratislava

Podpredsedovia:

- **pre aplikovanú štatistiku**
  Doc. Ing. Vladimír Úradníček, PhD., UMB Banská Bystrica
- **pre demografiu**
  Doc. RNDr. Branislav Bleha, PhD. PF UK Bratislava
- **pre matematickú štatistiku**
  Doc. RNDr. Rastislav Potocký, CSc., mim. prof., FMFI Univerzity Komenského Bratislava
- **pre štatistické riadenie kvality**
  Prof. RNDr. Michal Tkáč, CSc., EU Košice
- **pre štátnu štatistiku**
  Ing. František Bernadič, ŠÚ SR Bratislava
- **pre medinárodné styky**
  RNDr. Peter Mach, MDPT SR Bratislava

Hospodár:
Ing. Marek Radvanský, EÚ SAV Bratislava

Členovia sekretariátu:
Doc. RNDr. Karol Pastor, CSc., MFF Univerzity Komenského Bratislava
Ing. Iveta Stankovičová, PhD., FM UK Bratislava

Členovia výboru:

PhDr. Ľudmila Benkovičová, CSc.
Ing. František Bernadič
Ing. Ján Cuper.
PhDr. Ľudmila Ivančíková.
Ing. Anna Janusová.
Doc. RNDr. PaedDr. Stanislav Katina, PhD.
Prof. RNDr. Jozef Komorník, DrSc.
RNDr. Samuel Korónyi, PhD.
Doc. RNDr. Dagmar Kusendová., CSc.
Prof. RNDr. Jozef Mládek, DrSc.
Doc. RNDr .Oľga Nánásiová, CSc.
Doc. RNDr .Viliam Páleník, PhD.
Doc. RNDr. Karol Pastor, CSc.
Prof. RNDr. Rastislav Potocký, CSc.
Mgr. Michaela Potančoková, PhD.
Prof. RNDr. Beáta Stehlíková, CSc.
Mgr. Branislav Šprocha
Prof. RNDr. Anna Tirpáková, CSc.
Ing. Mária Vojtková, PhD.
Prof. RNDr. Gejza Wimmer, DrSc.

Revízna komisia:

Ing. Mikuláš Cár, CSc., NBS – predseda RK
Ing. Mária Kanderová, PhD., UMB Banská Bystrica
Doc. Mgr. Erik Šoltés, PhD., EU Bratislava

SLOVENSKÁ ŠTATISTICKÁ A DEMOGRAFICKÁ SPOLOČNOSŤ
MILETIČOVA 3, 82 467 BRATISLAVA

Schválené Valným zhromaždením SŠDS dňa 7.10.2010

STANOVY SPOLOČNOSTI

BRATISLAVA
2010

V záujme rozvoja štatistickej a demografickej vedy na území Slovenskej republiky pôsobí Slovenská štatistická a demografická spoločnosť, ako priama pokračovateľka Slovenskej demografickej a štatistickej spoločnosti pri SAV založenej 28.3.1968. Pod novým názvom pôsobí od 14.3.1990.

1

Postavenie, sídlo a pôsobnosť Spoločnosti

1. Slovenská štatistická a demografická spoločnosť (ďalej v texte stanov len Spoločnosť) je dobrovoľné, výberové združenie vedeckých a iných odborných pracovníkov v oblasti štatistiky a demografie, prípadne iných príbuzných disciplín.
2. Sídlom Spoločnosti je Bratislava.
3. Spoločnosť pôsobí na území Slovenskej republiky.
4. Spoločnosť vyvíja činnosť ako vedecká spoločnosť pri Slovenskej akadémii vied.
5. Spoločnosť v medzinárodných stykoch pôsobí pod názvom Slovak Statistical and Demographical Society.

2

Znak Spoločnosti

1. Znak Spoločnosti tvorí kruh o priemere 6 cm. Vo výške 1,5 cm od spodného okraja je vodorovná os, nad ktorou je zobrazená Gaussova krivka s maximom 3,8 cm. V poli pod Gaussovou krivkou je zobrazený znak Slovenskej republiky. Po obvode kruhu je nadpis SLOVENSKÁ ŠTATISTICKÁ A DEMOGRAFICKÁ SPOLOČNOSŤ. V anglickej verzii SLOVAK STATISTICAL AND DEMOGRAPHICAL SOCIETY.
2. Je dovolené znak úmerne zmenšovať alebo zväčšovať.
3. Grafické zobrazenie znaku Spoločnosti je uvedené v prílohe stanov Spoločnosti.

3

Poslanie a hlavné úlohy Spoločnosti

Poslaním Spoločnosti je najmä:

1. Rozvíjať štatistickú a demografickú vedu a jej spoločenské využitie na prospech Slovenskej republiky.
2. Rozširovať poznatky z oblasti štatistiky a demografie.
3. Organizovať konferencie, semináre, sympózia, prednášky, diskusie, odborné školenia a iné akcie.
4. Poskytovať pomoc svojim členom pri ich vedeckej a inej odbornej práci.
5. Zvyšovať odbornú úroveň svojich členov s osobitným zreteľom na mladých pracovníkov.
6. Prezentovať výsledky štatistickej a demografickej vedy v Slovenskej republike na domácich a medzinárodných fórach.
7. Rozvíjať vlastnú publikačnú a ediničnú činnosť.

4

Členstvo

1. Členovia Spoločnosti sú riadni, čestní a kolektívni.
2. Riadnym členom Spoločnosti sa môže stať vedecký a iný odborný pracovník v odboroch štatistiky alebo demografie alebo príbuzných disciplín.
3. Čestným členom Spoločnosti sa môže stať významný domáci alebo zahraničný vedecký alebo iný odborný pracovník, ktorý sa zaslúžil o rozvoj štatistickej alebo demografickej vedy. Čestné členstvo udeľuje Valné zhromaždenie na návrh Výboru.
4. Kolektívnym členom Spoločnosti sa môže stať právnická osoba (spravidla škola, vedecká alebo iná organizácia). Kolektívneho člena zastupuje jeden ňou poverený reprezentant, ktorý spĺňa podmienky pre riadneho člena.
5. Členstvo vzniká po zaplatení zápisného a členského poplatku dňom registrácie prihlášky Výborom Spoločnosti. Výšku zápisného a členského určuje Výbor Spoločnosti.

5

Práva a povinnosti členov

1. Člen Spoločnosti má právo:

a. voliť a byť volený do všetkých orgánov Spoločnosti;
b. podávať návrhy a hlasovať o návrhoch podaných na schôdzkach Spoločnosti;
c. byť informovaný a zúčastňovať sa na všetkých jej podujatiach;
d. prednostne získavať časopisy a iné publikácie, ktoré Spoločnosť vydáva.

2. Člen Spoločnosti má povinnosť:

a. zachovávať ustanovenia stanov a plniť platné rozhodnutia jej orgánov;
b. zúčastňovať sa na plnení úloh Spoločnosti a obhajovať jej záujmy;
c. platiť členské príspevky. Od platenia členských príspevkov sú oslobodení študenti a penzisti.

6

Zánik členstva

1. Členstvo v Spoločnosti zaniká úmrtím, vystúpením alebo zrušením.
2. Riadne členstvo môže byť zrušené najmä:

a. ak správanie alebo činnosť člena je v rozpore s povinnosťami, ktoré sú určené stanovami;
b. pre neplatenie členských príspevkov (napriek upomienkam) za obdobie dlhšie ako dva roky.

3. O zrušení členstva rozhoduje Výbor Spoločnosti.

7

Orgány Spoločnosti

1. Vedúcimi orgánmi Spoločnosti sú:

a. Valné zhromaždenie
b. Výbor.

2. Kontrolným orgánom Spoločnosti pre hospodársku činnosť je Revízna komisia.

8

Valné zhromaždenie

1. Najvyšším orgánom Spoločnosti je Valné zhromaždenie. Riadne Valné zhromaždenie zvoláva Výbor najmenej raz za štyri roky. Mimoriadne Valné zhromaždenie zvoláva Výbor z vlastnej iniciatívy, alebo na žiadosť aspoň jednej tretiny riadnych členov, a to najneskôr do jedného mesiaca po predložení žiadosti. Valného zhromaždenia má právo zúčastniť sa každý člen Spoločnosti.
2. Do právomoci Valného zhromaždenia patrí:

a. schvaľovať, upravovať a meniť stanovy Spoločnosti;
b. uznášať sa na hospodárskych opatreniach;
c. schvaľovať správy odstupujúceho Výboru a Revíznej komisie;
d. tajným hlasovaním voliť a odvolávať členov Výboru a Revíznej komisie;
e. rozhodovať o odvolaniach proti rozhodnutiu Výboru;
f. voliť čestných členov Spoločnosti;
g. rušiť čestné členstvo Spoločnosti;
h. uznášať sa o zrušení Spoločnosti.

3. Valné zhromaždenie riadi predseda alebo podpredseda alebo iný poverený člen Spoločnosti.

1. Valné zhromaždenie je schopné právoplatne sa uznášať, ak je prítomná najmenej polovica riadnych členov. Ak sa v určený čas nedostaví potrebný počet členov, po polhodinovej čakacej lehote sú uznášania schopní prítomní členovia.
2. Za prijatie uznesenia musí hlasovať nadpolovičná väčšina prítomných členov. Pri rovnosti hlasov rozhoduje hlas predsedu.

9

Výbor Spoločnosti

1. Výbor je výkonným orgánom Spoločnosti. Riadi činnosť Spoločnosti v období medzi Valnými zhromaždeniami.
2. Výbor plní uznesenia Valného zhromaždenia Spoločnosti.
3. Výbor sa skladá z predsedu, podpredsedov, vedeckého tajomníka, hospodára a ďalších členov Výboru. Z titulu svojej funkcie je členom Výboru predseda Štatistického úradu Slovenskej republiky. Valné zhromaždenie môže určiť ďalších členov Výboru z titulu ich funkcií vo významných štatistických alebo demografických pracoviskách. Z titulu svojich funkcií sú členmi Výboru predsedovia pobočiek a sekcií Spoločnosti. Počet členov Výboru určí Valné zhromaždenie s prihliadnutím na celkový počet členstva Spoločnosti. Výbor sa volí na dobu štyroch rokov. Členmi Výboru môžu byť len riadni členovia Spoločnosti.
4. Predsedu, podpredsedov, vedeckého tajomníka a hospodára volí Výbor Spoločnosti spomedzi svojich členov tajným hlasovaním.
5. Za členov Výboru môže Výbor kooptovať nových členov.
6. Schôdze Výboru Spoločnosti zvoláva predseda alebo vedecký tajomník podľa potreby, najmenej dvakrát do roka; prípadne do jedného týždňa po požiadaní aspoň tretiny členov Výboru.

7. Výbor je schopný uznášať sa za prítomnosti aspoň tretiny svojich členov. Ak sa v určený čas nedostaví potrebný počet členov, po polhodinovej čakacej lehote sú uznášania schopní prítomní členovia.
8. Za uznesenie Výboru sa považuje návrh, za ktorý hlasovala väčšina prítomných. V prípade verejného hlasovania pri rovnosti hlasov rozhoduje hlas predsedu.

10

Revízna komisia

1. Dozor nad hospodárskou činnosťou Spoločnosti vykonáva Revízna komisia.
2. Revíznu komisiu volí Valné zhromaždenie z členov Spoločnosti na funkčné obdobie zhodné s funkčným obdobím Výboru. Členovia Revíznej komisie nesmú byť členmi Výboru Spoločnosti. Revízna komisia volí spomedzi svojich členov svojho predsedu. Revízna komisia môže kooptovať nových členov z radov členov Spoločnosti.
3. Členovia Revíznej komisie sa môžu zúčastňovať s hlasom poradným na zasadnutiach Výboru Spoločnosti.
4. Revízna komisia podáva správy o svojej činnosti Valnému zhromaždeniu. Bez jej súhlasu nemôže byť prerokovaná účtovná uzávierka, ani prijaté uznesenie o použití prebytkov alebo úhrade straty.

11

Sekretariát Spoločnosti

1. Sekretariát Spoločnosti tvorí predseda, vedecký tajomník, podpredsedovia, hospodár, administratívny pracovník a prípadne ďalší členovia Spoločnosti, ktorých určí Výbor.
2. Úlohou Sekretariátu je výkon operatívnych činností Spoločnosti.
3. Práca Sekretariátu podlieha kontrole Výboru.

12

Pobočky

1. Výbor Spoločnosti zriaďuje teritoriálne pobočky.
2. Úlohou pobočiek je plniť úlohy Spoločnosti v okruhu svojej pôsobnosti.

13

Sekcie a odborné skupiny

1. Na lepšie zabezpečenie činnosti zriaďuje Spoločnosť sekcie alebo odborné skupiny, do ktorých sa členovia združujú podľa svojich záujmov a špecializácie.
2. Sekcie a odborné skupiny sa vytvárajú bez ohľadu na teritoriálne členenie Spoločnosti. Podľa potreby sa môžu organizovať aj v rámci pobočiek.
3. Podmienkou členstva v sekcii alebo odbornej skupine je členstvo v Spoločnosti. Člen môže pracovať v ľubovoľnom počte sekcií alebo odborných skupín.
4. Sekciu vedie podpredseda Spoločnosti.

14

Hmotné prostriedky a hospodárenie Spoločnosti

1. Hmotné prostriedky na činnosť Spoločnosti tvoria zápisné, členské príspevky, dotácie poskytované Slovenskou akadémiou vied alebo inými ustanovizňami a príjmy z vlastnej činnosti. Spoločnosť predkladá orgánom SAV výročnú správu o činnosti a hospodárení.
2. Výbor zostaví koncom každého roka plán činnosti a na jeho podklade finančný rozpočet na budúci rok. Okrem toho začiatkom roka zostaví účtovnú uzávierku za uplynulý rok.
3. Spoločnosť hospodári na základe schváleného plánu a rozpočtu na príslušný rok. Platby a príjmy Spoločnosti prebiehajú na účet Spoločnosti. V rámci svojej hospodárskej činnosti Spoločnosť organizuje vedecké, odborné a propagačné akcie s vložným resp. školným a ďalšími poplatkami, realizuje vydávanie a predaj publikácií, v svojich publikáciách a na svojich akciách realizuje propagačnú a reklamnú činnosť za úhradu, organizuje výskumnú, vedeckú, vzdelávaciu a odbornú činnosť za úhradu, realizuje poradenstvo na úseku štatistiky a demografie a príbuzných oblastí za úhradu, zabezpečuje iné činnosti zabezpečujúce príjem prostriedkov na činnosť a chod spoločnosti.
4. Za majetok Spoločnosti zodpovedá Výbor, ktorý správou majetku poverí jedného člena (hospodára).

15

Zastupovanie Spoločnosti

1. V mene Spoločnosti rokujú a podpisujú nasledovní štatutárni zástupcovia: predseda, vedecký tajomník, hospodár. K svojmu podpisu pripojujú názov Spoločnosti a označenie funkcie. Vo veciach hospodársko-finančných sa vyžadujú podpisy dvoch štatutárnych zástupcov Spoločnosti.
2. Výbor môže splnomocniť svojich jednotlivých členov, prípadne aj iné osoby, aby v rozsahu, ktorý určí Výbor zastupovali Spoločnosť a rokovali v jej mene. Osoby takto poverené podpisujú za Spoločnosť s dodatkom “v zastúpení”.

16

Spolupráca

1. Slovenská štatistická a demografická spoločnosť spolupracuje so spoločnosťami, ktoré majú podobné úlohy v Slovenskej republike a v zahraničí.
2. Spoločnosť spolupracuje pri plnení svojich úloh, predovšetkým so Štatistickým úradom Slovenskej republiky, s pracoviskami SAV, vysokými školami, inými vedeckovýskumnými pracoviskami a ďalšími právnickými osobami.

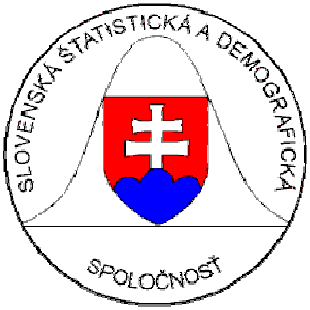
17

Zánik Spoločnosti

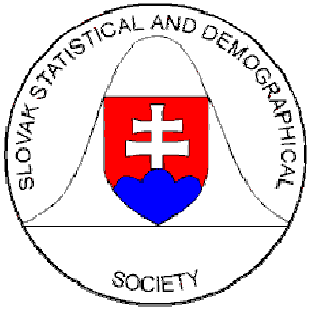
1. Spoločnosť zanikne, ak sa o jej zániku uznesie Valné zhromaždenie dvojtretinovou väčšinou hlasov prítomných členov. Valné zhromaždenie pre účely likvidácie práv a záväzkov Spoločnosti ustanoví likvidačný orgán.
2. O naložení s majetkom, ktorý zostane po úhrade jej záväzkov, rozhodne Valné zhromaždenie.

Príloha
Znak Slovenskej štatistickej a demografickej spoločnosti

a) slovenská verzia



b) anglická verzia



Schválené Valným zhromaždením Slovenskej štatistickej a demografickej spoločnosti dňa 7. 10. 2010.

Za správnosť:
Doc. Ing. Jozef Chajdiak, CSc.
Predseda SŠDS

OBSAH / CONTENT