INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 22, číslo 3–4, prosinec 2011

REGRESE JOINPOINT POMOCÍ PROGRAMU R

Jiří Anděl

Adresa: KPMS MFF UK, Sokolovská 83, 186 00, Praha 8

E-mail: andel@karlin.mff.cuni.cz

Abstrakt

Článek je věnován spojité lineární lomené regresi, která se nazývá joinpoint regression nebo také broken-line relationship. V poslední době se tato regrese používá poměrně často v medicíně. V programu R je obsažena knihovna segmented, pomocí níž lze výpočty provádět. V příspěvku je podrobně popsáno, jak se tato knihovna používá a jaká je interpretace získaných výsledků. Je upozorněno na to, že v některých případech výpočet selhává.

The paper describes joinpoint regression, which is also called broken-line relationship. This model is quite often used in medicine. The program R contains package segmented, which enables to calculate the joinpoint regression. It is described in detail how to use this package and how to interpret the results. It is remarked that in some cases the calculation fails.

1. Úvod

V posledních několika létech se začala hodně používat spojitá lineární lomená regresní funkce. Někteří autoři píší, že jde o „broken-line relationship“, častěji se však používá termín „joinpoint regression“. Časté je použití v medicíně, viz http://surveillance.cancer.gov/joinpoint/.

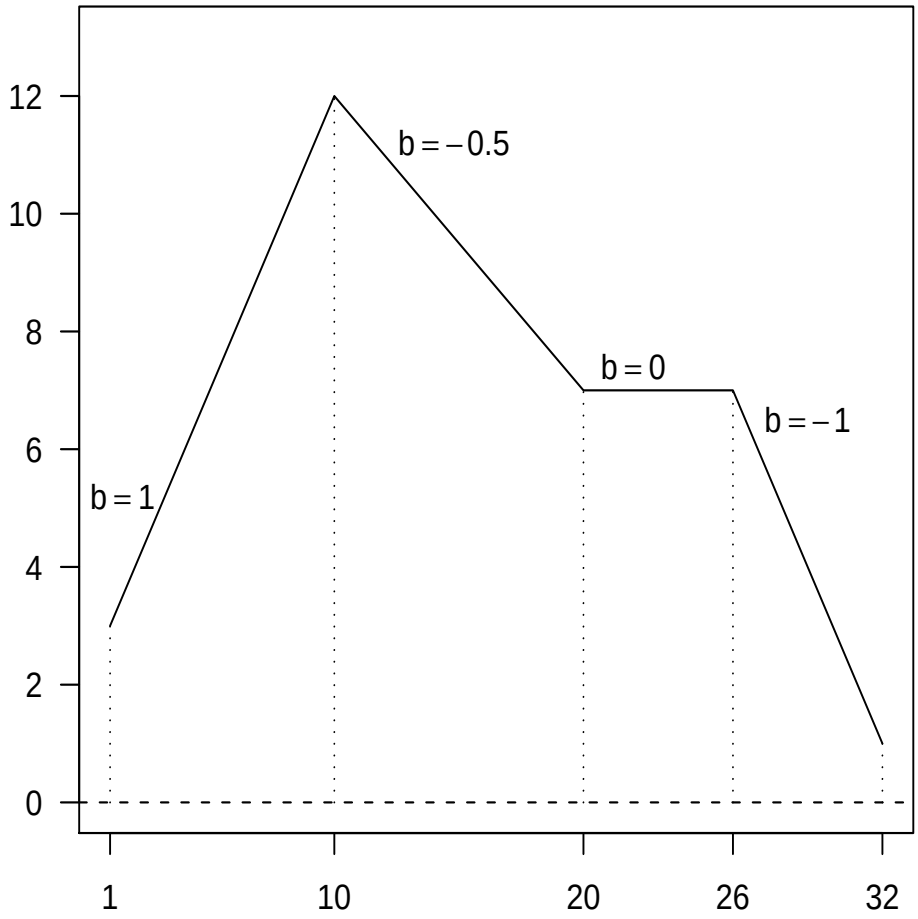
Najdou se aplikace i v jiných oblastech (např. při průzkumu citovanosti teoretických statistických článků v aplikovaných časopisech). Přitažlivá je nepochybně snadná interpretace vypočtených parametrů. To vynikne zejména v případě, kdybychom stejná data chtěli vyrovnávat polynomem nějakého vyššího stupně. Některé metody odhadu parametrů jsou založeny na práci Kim et al. (2000).

Po malém pátrání se dá zjistit, že program na výpočet lineární lomené regrese je obsažen i v programu R. Jeho stručný popis se najde v článku Muggeo (2008). V našem příspěvku popíšeme některé zkušenosti s prací s tímto programem. Budeme analyzovat simulovaná data, protože tam máme možnost porovnat odhady parametrů s jejich skutečnými hodnotami. Zde uvedeme i velmi podrobný program psaný v R, pomocí něhož byly výpočty i grafy získávány, protože si někteří kolegové stěžovali, že v minulém příspěvku (Anděl 2010) tomu tak nebylo.

Budeme se zabývat lineární lomenou regresní funkcí, která je znázorněna na obr. 1. Rovnice této funkce je

$$y = 2 + x - 1.5(x - 10)^+ + 0.5(0, x - 20)^+ - (x - 26)^+,$$

přičemž používáme známé označení $a^+ = \max(0, a)$.



Obrázek 1: Lomená regresní funkce

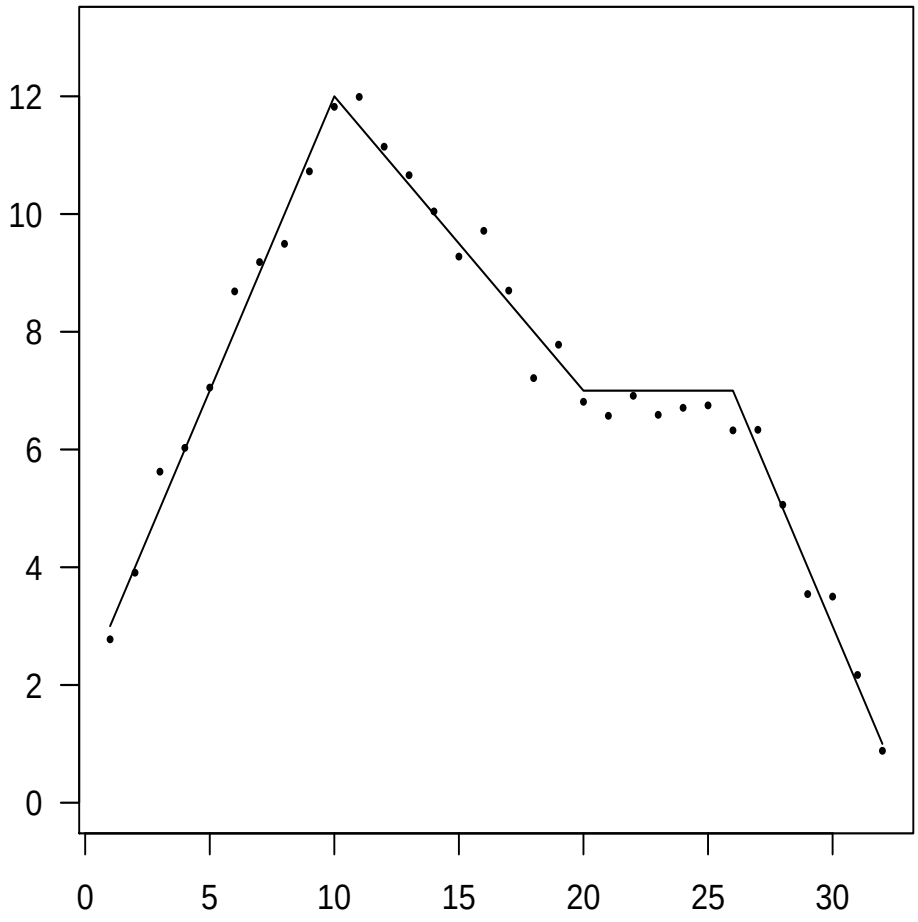
Obr. 1 byl vytvořen programem

```
x <- 1:32
yy <- 2 + x -1.5*pmax(0,x-10)+0.5*pmax(0,x-20)-1*pmax(0,x-26)
plot(x,yy, type="l", las=1, xlab="", ylab="", ylim=c(0,13),
     xaxt="n")
x0 <- c(1,10,20,26,32); y0 <- rep(0,5)
x1 <- x0; y1 <- c(3,12,7,7,1)
axis(1, at=x0, lab=x0)
segments(x0,y0,x1,y1, lty=3)
abline(h=0, lty=2)
```

```
text(1.5,5.2, expression(b==1))
text(14.8,11.2, expression(b==-0.5))
text(22,7.4, expression(b==0))
text(29,6.5, expression(b==-1))
```

Nyní vytvoříme simulovaná data. K výše uvedené regresní funkci přidáme nezávislé chyby s rozdělením $\mathsf{N}(0, 0.4)$. Výsledek je na obr. 2.

```
set.seed(123)
rnd <- rnorm(32,0,0.4)
y <- yy+rnd
plot(x,yy, type="l", las=1, xlab="", ylab="", ylim=c(0,13))
points(x,y,pch=16, cex=0.5)
```



Obrázek 2: Lomená regresní funkce a simulovaná data

2. Výpočet v programu R

K odhadu parametrů lineární lomené regresní funkce je k dispozici knihovna `segmented`. Je jí věnován článek Muggeo (2008) a popis jejích funkcí je samozřejmě obsažen v příslušném helpu.

Stručně připomeňme, že se nejprve vychází z obvyklé regrese vypočtené příkazem `lm` nebo `glm`. Dále `seg.Z` je vektor, v němž budou odhadovány body zlomu. Ve vektoru `psi` je počáteční odhad bodů zlomu. Je nutno podotknout, že tím je dán počet bodů zlomu. Program jejich polohu dál upřesňuje, ale jejich počet nemění. To by mohl uživatel dělat ručně např. využitím kritéria BIC. Výpočet se provede takto:

```
library(segmented)
dati <- data.frame(x,y)
out.lm <- lm(y~x, data=dati)
fit.seg<-segmented(out.lm,seg.Z=~x,psi=list(x=c(10,20,26)),
    control=seg.control(display=FALSE))
```

Nyní postupně vypíšeme a znázorníme vypočtené výsledky. Nejdřív budeme mít

```
slope(fit.seg)
$x
           Est. St.Err.  t value CI(95%).l CI(95%).u
slope1  0.96720 0.04069  23.7700    0.8832    1.0510
slope2 -0.55140 0.04069 -13.5500   -0.6354   -0.4674
slope3 -0.04592 0.08835  -0.5198   -0.2283    0.1364
slope4 -1.02800 0.08835 -11.6400   -1.2100   -0.8457
```

Výsledek je zřejmý. Ve sloupci `Est` máme odhadnuté směrnice jednotlivých úseků, obsah sloupců je uveden v záhlaví. Podrobnější výpis získáme takto:

```
summary.segmented(fit.seg)

        ***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = out.lm, seg.Z = ~x, psi = list(x = c(10, 20,
    26)), control = seg.control(display = FALSE))

Estimated Break-Point(s):
```

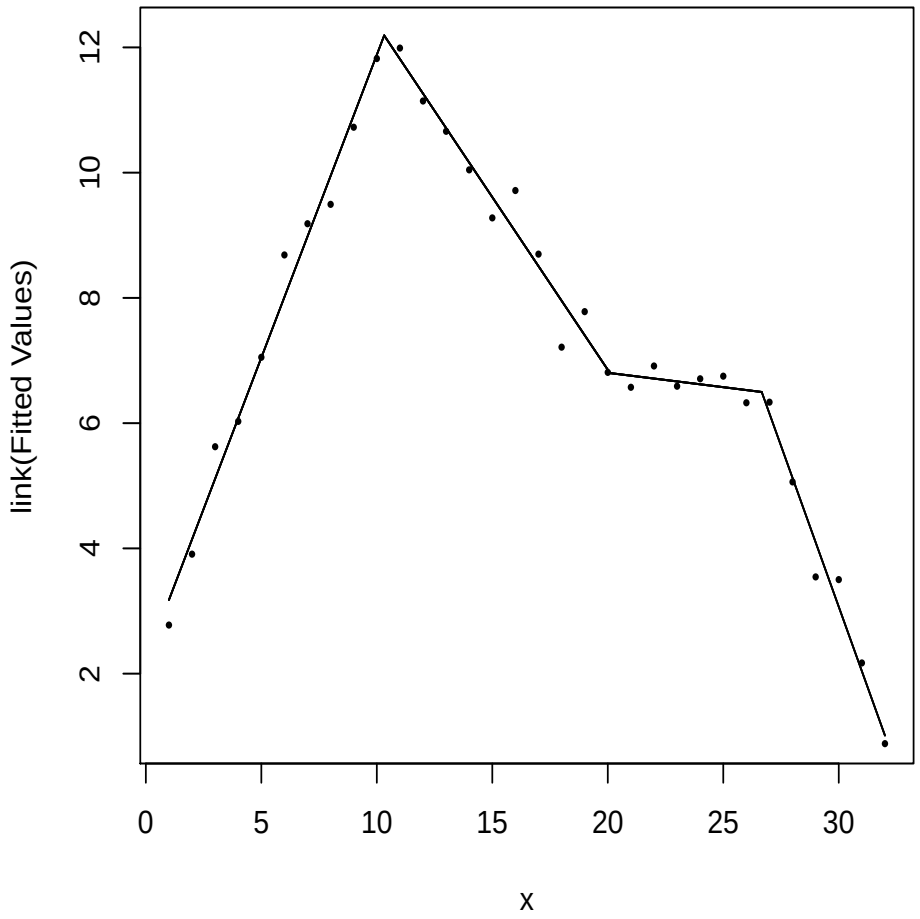
```
         Est. St.Err
psi1.x 10.32 0.2186
psi2.x 20.10 0.7959
psi3.x 26.66 0.4397

t value for the gap-variable(s) V:  1.059328e-15 -3.009219e-15
  2.553323e-15

Meaningful coefficients of the linear terms:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.21019    0.25248   8.754 6.18e-09 ***
x            0.96721    0.04069  23.770  < 2e-16 ***
U1.x        -1.51864    0.05755 -26.390       NA
U2.x         0.50550    0.09727   5.197       NA
U3.x        -0.98211    0.12495  -7.860       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3696 on 24 degrees of freedom
Multiple R-Squared: 0.9869,  Adjusted R-squared: 0.9831

Convergence attained in 2 iterations with relative change
  9.482222e-16
```

Veličiny `U1.x`, `U2.x` a `U3.x` popisují, o kolik se musí změnit směrnice v bodech zlomu proti směrnici v předchozím úseku. Graficky je výsledek znázorněn na obr. 3. Příslušné příkazy jsou:

```
plot.segmented(fit.seg)
points(x,y,pch=16, cex=0.5)
```

3. Některé technické poznámky

V jistých případech se stává, že iterační postup použitý ve funkci `segmented` nekonverguje. Tento případ může být signalizován tím, že program oznámí využití maximálního povoleného počtu iterací. Můžeme se o tom přesvědčit tím, že ve výpočtu použijeme `control=seg.control(display =TRUE)`. Tím dostaneme výpis jednotlivých kroků. Dalším varovným signálem jsou velké hodnoty veličin uvedených v `t value for the gap-variable(s) V`. Tento případ např. nastane, když zvýšíme rozptyl chybové složky, tedy když místo



Obrázek 3: Data a proložená regresní funkce

`rnd <- rnorm(32,0,0.4)` použijeme `rnd <- rnorm(32,0,0.6)`. Muggeo doporučuje zvolit menší krok `h`, např. použít `control=seg.control (display = FALSE), h=0.3`. Použijeme-li příkaz `rnd <- rnorm(32,0,0.6)` však ani tato rada nepomohla, a to ani v případě dalšího zmenšení kroku `h`.

Literatura

[1] Anděl J. (2010): Statistika a počítače, studenti a učitelé. *Inf. Bull. Čes. statist. spol.* **22**, č. 4, 8–16.

[2] Kim H. J., Fay M. P., Feuer E. J., Midthune D. N. (2000): Permutation tests for joinpoint regression with applications to cancer rates. *Stat. Med.* **19**, 335–351. Correction 2001, **20**, 655.

[3] Muggeo V. M. R. (2008): segmented: An R Package to Fit Regression Models with Broken-Line Relationships. *R News* 8(1), 20–25.

JAK NA ODHAD JOINPOINT REGRESE

Šárka Hudecová

Adresa: KPMS MFF UK, Sokolovská 83, 186 00, Praha 8

E-mail: `hudecova@karlin.mff.cuni.cz`

Abstrakt

V příspěvku[1] se zabýváme modelem joinpoint regrese, tj. modelem po částech lineární spojité závislosti. Tento model nachází zajímavá uplatnění v řadě oblastí. Pro jeho odhad byl vyvinut speciální software Joinpoint Regression Program, jehož metodologii a použití blíže popíšeme. Dále čtenáře seznámíme s knihovnou `segmented`, která je k dispozici v programu R. V obou případech uvedeme hlavní výhody a nevýhody daného softwaru a některé problémy, na něž může uživatel při práci narazit. Na několika datových souborech (simulovaných i reálných) provedeme praktické porovnání chování obou softwarů.

We deal with a joinpoint regression model, i.e. with a piecewise linear continuous regression. This model is common in many fields and finds various interesting applications. We give a description of the Joinpoint Regression Program, a special software developed for the estimation of the joinpoint regression models. The library `segmented` available in program R is presented as well. We provide a practical comparison of these two programs based on analyses of several data sets, simulated as well as real data. Advantages, disadvantages, and possible problems with estimation are discussed.

1. Úvod

Regrese joinpoint (v literatuře také *segmented regression*, *piecewise regression*, *broken line regression*) je model, v němž je závislost odezvy na vysvětlující proměnné popsána po částech lineární spojitou funkcí. Ta mění svou směrnici v několika obecně neznámých bodech zlomu (*transition points*, *break-points*, *change-points*, *joinpoints*). V některých praktických situacích takovýto model vyvstává zcela přirozeně z podstaty sledovaného problému, jinde ho lze s úspěchem použít k přibližnému popsání komplikovanější nelineární závislosti. Výhodou joinpoint regrese je zejména snadná interpretace parametrů, která nám umožňuje velice jednoduše popsat změny ve sledované závislosti (resp. sledovaném trendu). V mnohých aplikacích je také velmi

[1] Reakce na příspěvek prof. Anděla: *Regrese joinpoint s programem R*.

důležitá identifikace bodu zvratu, v němž dochází ke změně směrnice závislosti. Navíc s joinpoint regresní funkcí je možné pracovat také v případě složitějších modelů jako jsou zobecněné lineární modely (např. logistická regrese, loglineární model), Coxův model přežití a další.

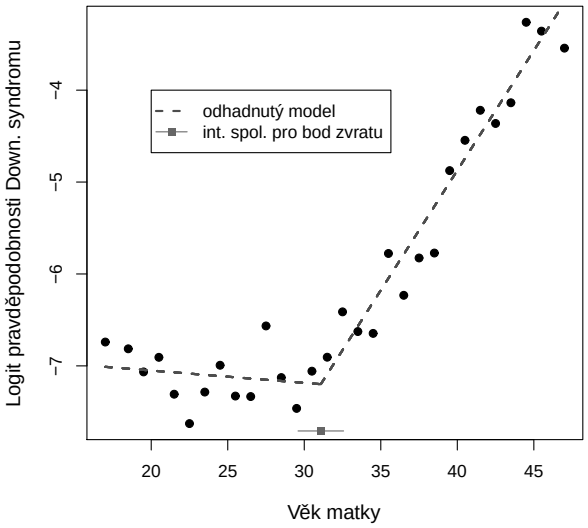
Je zřejmé, že joinpoint model je speciální případ regresního splinu, kdy po odhadované funkci vyžadujeme pouze spojitost. Avšak v joinpoint regresi mají body zvratu často velmi konkrétní (např. biologický nebo fyzikální) význam, jelikož v nich dochází ke strukturální změně sledované závislosti. Proto je nutné odhadu jejich počtu a polohy věnovat větší pozornost než v případě obecných splinů. Poznamenejme, že standardní metody založené na maximální věrohodnosti nelze pro joinpoint model automaticky použít, jelikož nejsou splněny klasické podmínky regularity, viz např. [2].

V následujícím textu nejdříve uvedeme několik vybraných zajímavých příkladů aplikací joinpoint regrese. Poté popíšeme, jak lze tento model odhadovat, a to jak pomocí speciálního softwaru Joinpoint Regression Program, tak v programu R pomocí knihovny `segmented`. Nakonec se pokusíme o stručné porovnání našich zkušeností s oběma softwary.

2. Vybrané příklady aplikací

Joinpoint regrese je aktuální téma, které nachází uplatnění v řadě oblastí, především v biostatistice, epidemiologii, biologii, chemii a dalších. Z konkrétních zajímavých aplikací jmenujme např. modelování pravděpodobnosti výskytu Downova syndromu u dítěte v závislosti na věku matky, viz [9] a náš obrázek 4, analýzu výskytu některých druhů rakoviny, či sledování výskytu některých infekčních chorob, viz [1]. Další konkrétní příklady využití v různých odvětvích jsou uvedeny např. v článku [8].

Využití joinpoint regrese pro sledování změn v trendu úmrtnosti a výskytu některých druhů rakoviny bylo hlavní motivací pro vznik metodologie navržené v práci [5], na jejímž základě byl vyvinut také speciální software Joinpoint Regression Program, o němž se blíže zmíníme v části 3. Autoři [5] ukazují jako příklad aplikaci joinpoint modelu na data úmrtnosti a výskytu rakoviny prostaty v čase. Je obecně známo, že výskyt tohoto typu rakoviny zaznamenal v posledních letech výrazné změny, a to především z důvodu zavedení screeningového testu PSA (prostate specific antigen). Zavedení jakéhokoliv obdobného testu má vždy za následek nejdříve rapidní zvýšení výskytu choroby (o případy, které by jinak byly objeveny mnohem později) časem následované poklesem. Joinpoint model umožňuje detekovat tyto právě popsané změny v trendu a následně pak lépe porozumět některým nepozorovatelným charakteristikám PSA testu jako jsou např. tzv. *lead time* (o jaký čas dříve



Obrázek 4: Logit pravděpodobnosti výskytu Downova syndromu dítěte v závislosti na věku matky. Odhadnutý model byt spočten v programu R funkcí `segmented`.

je choroba odhalena pomocí tohoto testu než by byla objevena standardními technikami) a *overdiagnosis* (irelevantní diagnóza „choroby“, která by nikdy nevytvořila symptomy a nevedla by k úmrtí). Joinpoint model odhadnutý pro data úmrtnosti umožňuje naopak popsat benefity PSA screeningu.

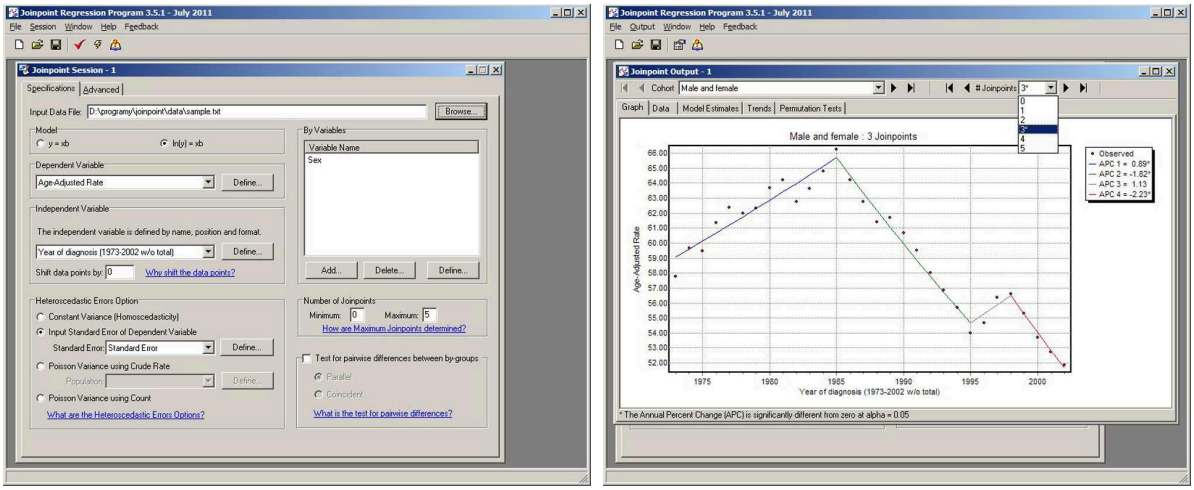
Zajímavé využití joinpoint regrese lze nalézt také v článku [10], kde je spojitá po částech lineární funkce použita k modelování chování počtu citací statistických článků v závislosti na počtu let od publikace. Nejprve jsou analyzovány rozdílnosti v trendu počtu citací v čase pro přední statistické časopisy ve srovnání s časopisem Journal of Clinical Oncology (JCO), viz také obrázek 9. Dále je joinpoint model aplikován přímo na jednotlivé nejvíce citované statistické články a jsou zachycené různé trendy chování v čase. Zájemce o tento problém odkážeme na `www.beststatisticalpractices.org`, kde je k dispozici jak článek [10], tak i aktualizovaný seznam „nejvýznamnějších“ (ve smyslu nejvyššího počtu tzv. aplikovaných citací) statistických článků.

3. Software Joinpoint Regression Program

Kim a kol. ve svém článku [5] navrhují postup, na jehož základě lze určit počet bodů zvratu a odhadnout parametry joinpoint modelu, včetně polohy bodů zvratu. Metoda je navržena jak pro standardní situaci, kdy jsou

regresní chyby nekorelované náhodné veličiny s konstantním rozptylem, tak i pro případ heteroskedastických a autokorelovaných chyb. Tím je umožněna mj. práce s poissonovskými odezvami, což je výhodné zejména pro modelování výskytu nějakého jevu (např. choroby) nebo úmrtnosti v čase.

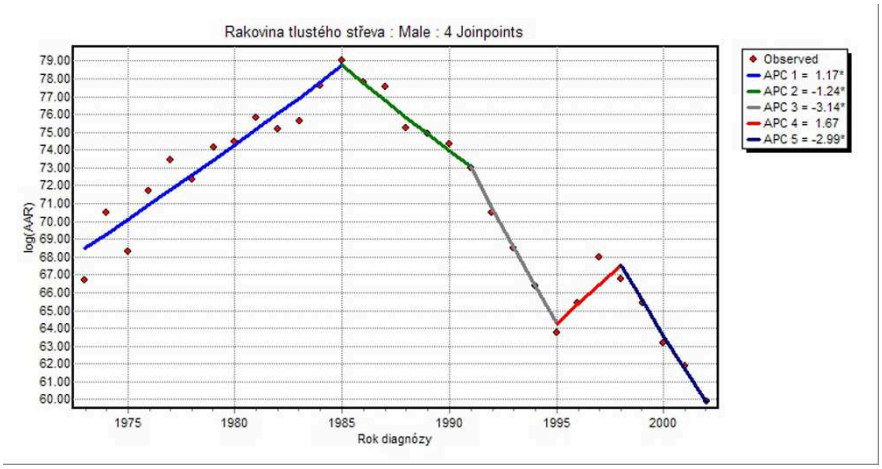
Identifikace počtu bodů zvratu je založena na sekvenci několika permutačních testů založených na modifikaci klasické F-statistiky. Tyto testy dosahují předepsané hladiny spolehlivosti asymptoticky a jejich p-hodnoty jsou spočteny pomocí Monte Carlo metody. Mnohonásobné testování je ošetřeno Bonferroniho korekcí. Parametry modelu jsou odhadovány metodou *grid search* navrženou v článku [7]. Body zvratu jsou nalezeny iteračně (prohledáváním mřížky) a odhad ostatních regresních parametrů je proveden metodou nejmenších čtverců, resp. vážených nejmenších čtverců.



Obrázek 5: Joinpoint Regression Program: vstupní dialog (vlevo) a výstupní dialog (vpravo).

Na základě výše popsané metodiky byl vyvinut speciální software Joinpoint Regression Program (v dalším jen JRP) pro odhad joinpoint modelu, viz [4]. Tento program je k dispozici zdarma po zaregistrování se na webové stránce `http://surveillance.cancer.gov/joinpoint/`. V současné době (prosinec 2011) je k dispozici verze 3.5.2, která již zaznamenala několik vylepšení a rozšíření oproti článku [5]. Kromě permutačního testu je možné vybrat finální model na základě BIC nebo modifikovaného BIC kritéria. Odhad parametrů lze provádět jak metodou grid search, tak i pomocí Hudsonovy metody, viz [3]. Dále je možné provádět porovnání joinpoint modelu pro dvě skupiny dat. Konkrétně je k dispozici test paralelnosti a identity založený na metodě popsané v [6]. Upravena je také korekce pro vícenásobné testování, které je nyní méně konzervativní než původně použitá Bonferroniho korekce.

JRP běží pod operačním systémem Windows (Windows 95 a novější). Je uživatelsky velice pohodlný a přehledný, viz obrázek 5. Uživatel zvolí „zaklknutím“, zda chce modelovat přímo závisle proměnnou nebo její logaritmus, a vybere, zda je uvažován model s konstantním rozptylem nebo model heteroskedastický. Zde uživatel buď sám specifikuje směrodatnou odchylku, nebo vybere model s předpokládaným Poissonovým rozdělením. Následně vybere minimální a maximální počet bodů zvratu, který má být uvažován, metodu odhadu parametrů (grid search nebo Hudsonovu), kritérium pro výběr nejlepšího modelu (permutační test, BIC, modifikované BIC) a počet Monte Carlo simulací pro výpočet p-hodnoty permutačního testu. Defaultní nastavení je grid search a permutační test založený na 4499 simulacích. Kromě toho je možné odhadovat model s autokorelovanými chybami a regulovat některé další parametry.



Obrázek 6: Joinpoint model pro závislost výskytu rakoviny tlustého střeva u mužů v průběhu let.

Při odhadování modelu s mnoha pozorováními a vyšším počtem bodů zvratu se uživatel musí připravit na to, že výpočet nějakou dobu potrvá, na což ho ovšem program slušně upozorní. Výstup z programu je opět uživatelsky velmi pohodlný a přehledný, viz obrázek 5. Umožňuje prohlédnout si výsledky (graf a tabulky výsledků) pro všechny uvažované počty bodů zvratu. Všechny části výstupu (graf, vyrovnané hodnoty, odhady parametrů i výsledky jednotlivých permutačních testů) je možné exportovat. Příklad toho, jak vypadá graf exportovaný z JRP vložený do LaTeXu je uveden na obrázku 6.

Software JRP má dvě zásadní nevýhody. První je, že jej lze využít pouze pro odhad lineárního nebo loglineárního modelu (nelze odhadovat logistickou regresi, regresní model přežití aj.). Druhou, zásadnější, nevýhodou je,

že umožňuje pracovat pouze s modelem s jednou nezávisle proměnnou, což může být někdy poněkud limitující.

4. Knihovna `segmented` v programu R

V programu R je pro odhad po částech lineární spojité regresní funkce k dispozici knihovna `segmented`, jejíž popis lze nalézt v [9] a jejíž metodika je založena na článku [8]. Knihovna umožňuje odhadovat joinpoint model pro zobecněné lineární modely, přičemž je možné pracovat s více vysvětlujícími proměnnými.

Na rozdíl od algoritmu JRP programu není odhad modelu proveden metodou grid search, ale jiným „trikovým“ iteračním postupem. Uvažujme, že chceme odhadnout joinpoint model, který má zlom v bodě ψ, ve kterém se směrnice mění z β_1 na $\beta_2 = \beta_1 + \delta_1$, tj. model

$$y = \beta_0 + \beta_1 x + \delta_1 (x - \psi)^+. \tag{1}$$

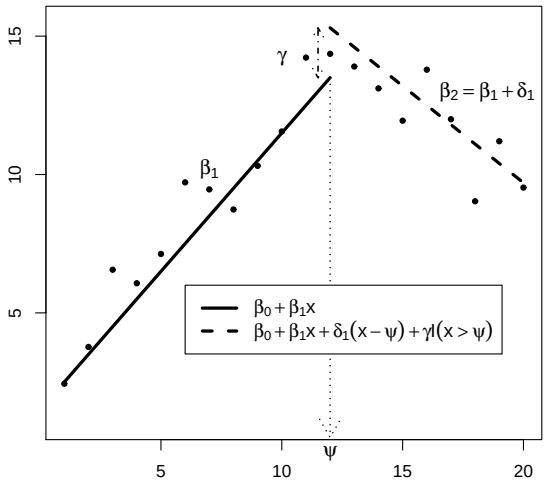
V práci [8] je ukázáno, že jestliže $\tilde{\psi}$ je počáteční odhad bodu zvratu ψ, pak model (1) je možné odhadnout iterativním odhadováním následujícího lineárního modelu

$$y = \beta_0 + \beta_1 x + \delta_1 (x - \tilde{\psi}) + \gamma \mathrm{I}(x > \tilde{\psi}), \tag{2}$$

kde $\mathrm{I}(\cdot)$ je identifikátor a γ je parametr, který měří nespojitost v bodě zvratu a pomocí něhož je přepočítáván odhad $\tilde{\psi}$, viz obrázek 7. Jestliže algoritmus konverguje, pak by výsledná regresní funkce měla být spojitá, tj. $\hat{\gamma} \approx 0$.

Knihovna `segmented` se nezabývá odhadem počtu bodů zvratu. Obsahuje sice test, který umožňuje pro zobecněný lineární model testovat, zda nastává změna ve směrnici závislosti či nikoliv (`davies.test`), ale [9] upozorňuje, že tento test není užitečný pro výběr vhodného počtu bodů zvratu a doporučuje použít spíše BIC kritérium nebo apriorní znalost problému.

Bohužel, věrohodnostní funkce v joinpoint modelu nemusí být konkávní, takže algoritmus nemusí nalézt globální maximum. Dále, navržený algoritmus v každém kroku pouze aproximuje skutečný model (1), což může také způsobovat určité problémy při odhadování. V praxi se proto doporučuje spustit program pro několik různých počátečních hodnot parametru ψ. Dostaneme-li různé hodnoty odhadu bodu zvratu, je možné požádat program o výpis věrohodnosti odpovídající jednotlivým modelům a „ručně“ vybrat ten nejvhodnější. Samozřejmě, čím výraznější je skutečná změna ve směrnici, tím menší je význam počáteční volby $\tilde{\psi}$. Autor metody doporučuje volit počáteční hodnotu bodu zvratu na základě posouzení grafického znázornění sledované závislosti.



Obrázek 7: Model, pomocí něhož je iterativně odhadnuta joinpoint regrese funkcí `segmented`.

Dalším problémem je situace, kdy algoritmus nekonverguje. Většinou nás na to program upozorní varovnou hláškou (byl dosažen maximální počet iteračních kroků). Navýšení počtu povolených iterací většinou problém nevyřeší, jelikož tato situace nastává často v případě, kdy minimalizovaná funkce (reziduální součet čtverců) alternuje mezi dvěma různými hodnotami. Jednou možností nápravy je zmenšení přírůstku, který se mezi jednotlivými kroky připočítává k dosavadnímu odhadu bodu zvratu. To ale bohužel problém dost často neřeší a konvergence není dosažena ani po této změně nastavení. V takovém případě program sice jakýsi odhad poskytne, ale jedná se o nespojitou funkci.

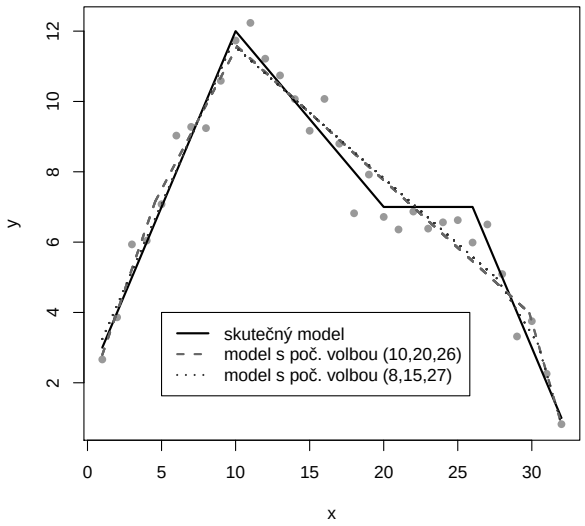
Autor [9] nabádá, že jestliže konvergence není dosažena automaticky nebo v případě, kdy obdržíme různé výsledky pro různá počáteční nastavení parametru ψ, může být parametrizace joinpoint modelem pro daná data diskutabilní. Doporučuje pak test přítomnosti bodu zvratu nebo posouzení BIC kritéria.

5. Porovnání

Některé rozlišnosti softwaru JRP a balíku `segmented` již byly popsány výše. Připomeňme, že JRP oproti `segmented` umožňuje testovat počet bodů zvratu a práci s autokorelovanými chybami. Naproti tomu `segmented` je schopen

pracovat s obecnějšími modely (zobecněné lineární modely, Coxův model přežití apod.) a s více než jednou proměnnou. Metody dohadů parametrů se taktéž liší, takže většinou nedostaneme úplně identické výsledky. JRP navíc defaultně předpokládá body zvratu v některé z x-ových souřadnic naměřených dat.

Nyní uvedeme porovnání chování obou softwarů na několika konkrétních datových souborech.



Obrázek 8: Dva „špatné“ modely pro simulovaná data se směrodatnou odchylkou 0.6 v porovnání se skutečným modelem.

5.1. Analýza simulovaných dat

V článku prof. Anděla byl proveden odhad joinpoint modelu v programu R funkcí `segmented` pro simulovaná data. Jak již bylo uvedeno, pro model s chybami se směrodatnou odchylkou 0.4 probíhá všechno bez problému, avšak pro případ chyb se směrodatnou odchylkou 0.6 již algoritmus nekonverguje. Zvolení menšího kroku nebo nastavení vyššího počtu iterací problém neřeší. Měníme-li dostatečně vytrvale startovací hodnoty pro joinpoint body zvratu, pak pro volbu $(11, 17, 29)$ lze dosáhnout konvergence k poněkud nesprávnému modelu, který má zvraty v bodech 4.6, 10.1 a 29.8. Pro nastavení $(8, 15, 27)$ zase dostaneme model se zvraty v bodech 9.8, 27.6 a 30.4. Z porovnání věrohodnostní funkce bychom mohli dojít k závěru, že první („špatný“) model je o něco lepší než druhý zmíněný („špatný“) model. Nicméně, oba se poměrně

Estimated Joinpoints			
Joinpoint	Estimate	Lower CI	Upper CI
1	10	9	12
2	20	17	23
3	27	24	28

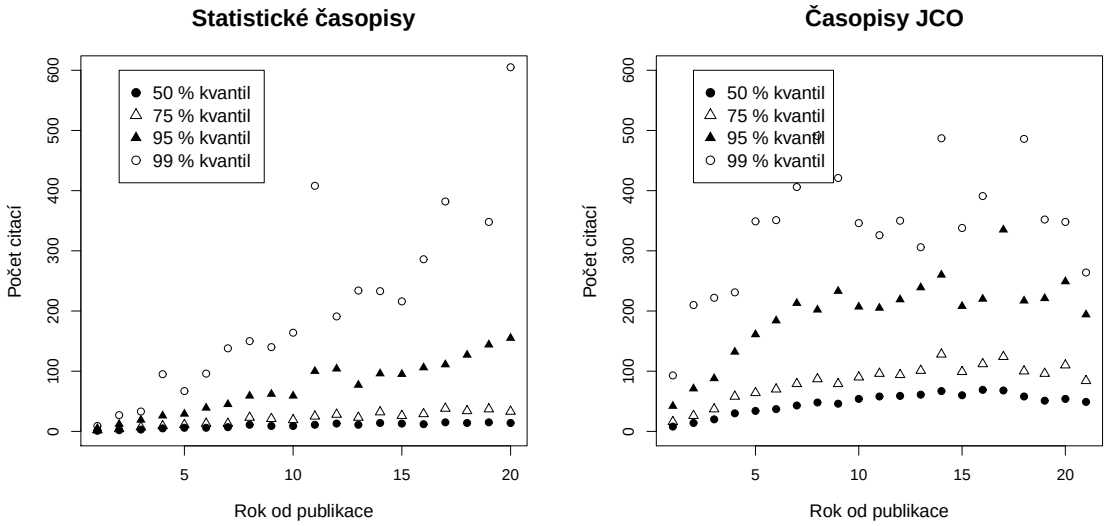
General Parameterization				
Parameter	Estimate	Standard Error	Z	Prob$> \|t\|$
Intercept 1	2,089	0,284	7,366	0,000
Intercept 2	17,225	0,767	22,455	0,000
Intercept 3	8,931	2,199	4,062	0,001
Intercept 4	34,877	3,708	9,407	0,000
Slope 1	1,000	0,050	19,848	0,000
Slope 2	−0,513	0,050	−10,185	0,000
Slope 3	−0,099	0,093	−1,057	0,303
Slope 4	−1,060	0,123	−8,583	0,000

Tabulka 1: Odhady bodů zvratu a parametrů odhadnuté v JRP pro simulovaná data se směrodatnou odchylkou 0.4.

dost liší od skutečného modelu, viz obrázek 8. Pro vyšší hodnoty směrodatné odchylky (0.8 a 1) je situace podobná až horší.

Stejná data jsme analyzovali také v programu JRP. Pro pozorování se směrodatnou odchylkou 0.4 dostáváme zanedbatelné rozdíly v bodových odhadech polohy bodů zvratu jakožto i v odhadech směrnice (např. 0.97 vs 1.00 v případě první směrnice). Výsledky lze tedy považovat za srovnatelné. Pro možnost hlubšího srovnání uvádíme některé z výstupů z JPR v tabulce 1. Pro model se směrodatnou odchylkou 0.6 JRP odhadne tři body zvratu (10, 21 a 27), bodové odhady směrnic jsou $1, -0.51, -0.05, -1.05$. JRP dále „zvládne“ i situaci, kde je směrodatná odchylka zvolena jako 0.8. Pro tento případ také vybere jako nejvhodnější model se třemi zvraty (v bodech 11, 20 a 27), přičemž odhady směrnic teď jsou $0.89, -0.65, -0.06, -1.01$. Pro rozptyl roven 1 už JRP shledává jako nejvhodnější model pouze s jedním joinpointem, a to v bodě 10. Odhaduje pak změnu směrnice z 0.95 na -0.43.

Na základě těchto výsledků lze konstatovat, že pro naše simulovaná data dává program JRP lepší výsledky než knihovna `segmented` v R.



Obrázek 9: Některé vyběrové kvantily počtu citací pro přední statistické časopisy a pro články z časopisu Journal of Clinical Oncology (JCO). Můžeme pozorovat dosti rozdílné trendy v čase.

5.2. Analýza počtu citací statistických článků

Již v části 2. jsme uvedli, že počet citací statistických článků byl analyzován pomocí joinpoint modelu v článku [10]. Grafy některých výběrových percentilů jsou uvedeny na obrázku 7. Pro všechny kvantily JCO je navržen model s jedním bodem zvratu, přičemž směrnice je poměrně vysoká pro první roky po publikaci (5 až 8 let) a výrazně nižší (až nulová) v dalších letech. Pro statistické časopisy je navržen model s jedním bodem zvratu pro 50% a 99% kvantil, zatímco pro 75% a 95% kvantil se zdá být závislost lineární.[2]

Poznamenejme, že na základě našich výsledků se zdá, že v původní analýze v [10] je menší nesrovnalost, jelikož počet „potřebných“ bodů zvratu byl zřejmě určen pomocí softwaru JRP na základě homoskedastického modelu, zatímco finální model je odhadnut pomocí vážených nejmenších čtverců a odpovídá tedy modelu heteroskedastickému. Výsledky jsou rozdílné zejména pro 99% kvantil, kde pro případ heteroskedastického modelu vychází jednoduchá lineární závislost jako dostačující k popisu chování počtu citací.

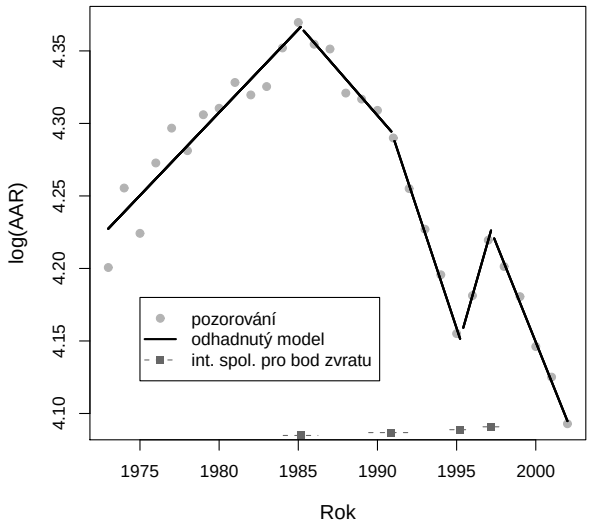
Srovnejme výsledky pro tato data v JRP s výsledky funkce `segmented`. V obou případech byl uvažován heteroskedastický lineární model. V R jsme

[2]Pro analýzu byla použita pouze data z let 1 až 16 pro JCO a data z let 1 až 19 pro statistické časopisy. Navíc, pozorování 11 bylo v případě 97.5% (není znázorněno na našem obrázku) a 99% kvantilu považováno za odlehlé a z analýzy vyřazeno.

navíc testovali, zda dochází ke změně ve směrnici lineárního modelu pomocí funkce `davies.test`.

Pro citace JCO dává R i JRP přibližně stejné výsledky. Pro statistické časopisy je situace složitější. Pro případ 50% kvantilu má `segmented` problém model odhadnout, přestože `davies.test` naznačuje statisticky významnou změnu směrnice. V případě 75% kvantilu podle JRP i podle R shodně neexistuje bod zvratu. Výsledky obou softwarů jsou odlišné pro 97.5% kvantil a 99% kvantil. Pro 97.5% kvantil dává R dosti přesný odhad bodu zvratu, zatímco JRP uvažuje jednoduchou lineární funkci. Pro 99% kvantil R indikuje významnou změnu trendu, ale funkce `segmented` takový model neumí odhadnout.

Z těchto výsledků lze usuzovat, že chování počtu citací v časopise JCO lze zřejmě popsat lineární funkcí s jedním bodem zlomu. Zde dávají oba softwary srovnatelné výsledky. Naopak, použití joinpoint modelu pro percentily počtu citací článků ze statistických časopisů je zřejmě trochu pochybné. Zejména výsledky pro vyšší kvantily (97.5% a 99%) uvedené v [10] a následné interpretace mohou být proto možná i mírně zavádějící.



Obrázek 10: Odhadnutý nespojitý model z programu R pro data výskytu rakoviny tlustého střeva u mužů.

5.3. Analýza výskytu rakoviny tlustého střeva

Na internetových stránkách softwaru JRP, viz [4], jsou k dispozici data o výskytu rakoviny tlustého střeva v USA v letech 1973–2002 pocházející z databáze SEER (Surveillance Epidemiology and End Results). Ke každému roku je k dispozici počet případů, velikost populace, výskyt choroby měřený pomocí AAR (*age adjusted rate*) a spočtená směrodatná chyba (spočteno v SEER), která se využívá pro výpočet vah v heteroskedastickém modelu. Data jsou k dispozici jak zvlášť pro muže a ženy, tak i bez rozlišení. Je uvažován heteroskedastický model závislosti logaritmu AAR na roku diagnózy s nekorelovanými chybami. Porovnejme výsledky JRP a `segmented` pro soubor mužů.

Vybíráme-li mezi modely s nula až pěti body zvratu, vybere JRP jako nejlepší model se čtyřmi body zvratu. Výsledné proložení je uvedeno na obrázku 6. Konkrétní odhady parametrů zde uvádět nebudeme.

Stejný model se čtyřmi body zvratu se pokusíme odhadnout v programu R funkcí `segmented`. Zadáme proto stejné váhy v JRP a zvolíme počáteční hodnoty bodů zvratu např. jako 1985, 1989, 1994 a 1998. Algoritmus v tomto případě konverguje ve čtyřech krocích a všechno se zdá být v pořádku, viz výstup uvedený v tabulce 2. Když si ovšem necháme vykreslit graf odhadnuté funkce, zjistíme, že je nespojitá. To je patrné z obrázku 10. Navíc lze spočítat, že odhadnuté body zvratu skutečně nejsou průsečíky jednotlivých odhadnutých přímek. Takový výsledek lze jen stěží považovat za uspokojivý. Poznamenejme ještě, že stejný problém nastává i pro model s méně body zvratu.

6. Závěr

V našem příspěvku jsme se pokusili podat stručný přehled toho, k čemu joinpoint model v praxi slouží a jak jej lze odhadovat. Provedli jsme porovnání softwaru [4] a knihovny `segmented` z programu R. Nutno podotknout, že z tohoto souboje nevychází ani jeden z nich jako jasný vítěz. Jak jsme již zmínili, JRP umožňuje pracovat pouze s poměrně jednoduchými modely s jedinou nezávisle proměnnou. Naproti tomu `segmented` dává v některých případech spíše neuspokojivé výsledky, viz např. odstavec 5.3.

Pro joinpoint logistickou regresi lze využít v programu R také speciální balík `ljr`, který je založen na článku [1]. Tato práce vychází z metody navržené v [5], přičemž nabízí alternativní způsob odhadu parametrů (pomocí podmíněné věrohodnostní funkce) a několik zobecnění. Je tedy možné, že se časem v R objeví knihovna vycházející přímo z [5] a používající stejné

```
***Regression Model with Segmented Relationship(s)***

Call:
segmented.lm(obj = mod1, seg.Z = ~year, psi = list(year = c(10,
    16, 21, 25) + 1973), control = seg.control(display = T, it.max = 20,
    h = 0.001))

Estimated Break-Point(s):
           Est. St.Err
psi1.year 1985 0.5260
psi2.year 1991 0.6866
psi3.year 1995 0.3139
psi4.year 1997 0.2435

t value for the gap-variable(s) V: 1.1032e-13 1.6078e-13 2.6054e-13
                                                                2.3227e-13

Meaningful coefficients of the linear terms:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.835e+01  1.816e+00 -10.108 2.64e-09 ***
year         1.145e-02  9.174e-04  12.477 6.81e-11 ***
U1.year     -2.400e-02  3.559e-03  -6.745       NA
U2.year     -2.035e-02  4.840e-03  -4.205       NA
U3.year      7.115e-02  1.513e-02   4.701       NA
U4.year     -6.547e-02  1.510e-02  -4.336       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.837 on 20 degrees of freedom
Multiple R-Squared: 0.9859,  Adjusted R-squared: 0.9795

Convergence attained in 4 iterations with relative change -3.496302e-15
```

Tabulka 2: Výsledek funkce `summary` pro joinpoint model se čtyřmi body zvratu pro data výskytu rakoviny tlustého střeva u mužů.

algoritmy jako software JRP. Na závěr ještě podotkněme, že v programu R je k dispozici také knihovna `strucchange`, který se zabývá tzv. modely se strukturálními změnami (aplikovanými především v ekonomii a ekonometrii). Zde se podobně předpokládá, že v určitých bodech zvratu dochází ke změně směrnice, ale není vyžadována spojitost regresní funkce, což situaci (odhad a inferenci) značně usnadňuje a jedná se tedy o zcela jiný problém.

Literatura

[1] Czajkowski M., Gill R. a Rempala G. (2008) Model selection in logistic joinpoint regression with applications to analyzing cohort mortality patterns. Stat. Med. **27**, 1508 – 1526.

[2] Feder P. I. (1975) On asymptotic distribution theory in segmented regression problems. Ann. Statist. **3**, 49 – 83.

[3] Hudson D. (1966) Fitting segmented curves whose join points have to be estimated. J. Amer. Statist. Assoc. **61**, 1097 – 1129.

[4] Joinpoint Regression Program (2011), Statistical Methodology and Applications Branch and Data Modeling Branch, Surveillance Research Program National Cancer Institute, version 3.5 — April 2011.
`http://surveillance.cancer.gov/joinpoint/`

[5] Kim H. J. a kol. (2000) Permutation tests for joinpoint regression with applications to cancer rates. Stat. Med. **19**, 335 – 351. Correction: Stat. Med. 20, 2001, 655.

[6] Kim a kol. (2004) Comparability of Segmented Line Regression Models. Biometrics **60**, 1005 – 1014.

[7] Lerman P. M. (1980) Fitting segmented regression models by grid search. Appl. Statist. **29**, 77 – 84.

[8] Muggeo V. M. R.(2003) Estimating regression models with unknown break-points. Stat. Med. **22**, 3055 – 3071.

[9] Muggeo V. M. R.(2008) Segmented: An R Package to fit regression models with broken-line relationships. R News **8**, 20 – 25.

[10] Schell M. J. (2010) Identifying key statistical papers from 1985 to 2002 using citation data for applied biostatisticians.
Amer. Statist. **64**, 310 – 317.

NĚKTERÉ METODY DATA MININGU, ZEJMÉNA PRO APLIKACE V KLINICKÉM ROZHODOVÁNÍ

Jan Kalina

Adresa: Ústav informatiky AV ČR, v.v.i., Pod Vodárenskou věží 2, 182 07 Praha 8

E-mail: kalina@euromise.cz

Abstrakt

Článek shrnuje obecné principy data miningu a popisuje roli data miningu v procesu medicínského rozhodování, zejména v kontextu systémů pro podporu rozhodování. Detailně popisuje populární algoritmus pro extrakci vzorů pro kategoriální data, přičemž jej rozebírá ze statistického hlediska. Nakonec zkoumá vztah mezi neuronovými sítěmi a logistickou regresí, které představují dvě nejčastěji používané data miningové metody v biomedicínských aplikacích.

The paper overviews general principles of data mining and discusses the role of data mining in the process of medical decision making, particularly in the context of decision support systems. We examine a popular algorithm for pattern discovery in categorical data and discuss it from the point of view of classical statistics. Further we investigate the relationship between neural networks and logistic regression, which are two most popular data mining methods in biomedical applications.

Klíčová slova: Systémy pro podporu rozhodování, analýza asociací, klasifikační analýza, neuronová síť.

Keywords: Decision support system, association analysis, classification analysis, neural network.

1. Principy data miningu

Data mining (dolování dat, vytěžování dat) lze charakterizovat jako proces extrakce informace z velkých datových souborů, který vede k odhalení a prozkoumání systematických vztahů mezi veličinami. Mezi běžně používané data miningové metody patří rozhodovací stromy, analýza asociací, shluková analýza, neuronové sítě, logistická regrese a další [9]. Jednotlivé metody se převážně zařazují do informatiky, zatímco v klasické statistice nevzbudily pozitivní zájem, přestože se jim věnuje velká pozornost v seriózní odborné literatuře [6]. Naproti tomu v data miningu nejsou oblíbeny některé klasické

statistické metody jako například lineární diskriminační analýza. [16] popsal data mining jako „paběrkování na datových smetištích“, nad kterým „fundamentalističtí statistikové trochu ohrnují nos“. Také [4] připouští, že data mining ještě dnes zní některým statistikům jako sprosté slovo, případně alespoň jako naprosto nezajímavá metodologie.

[9] popsal data mining jako jeden z kroků procesu objevování znalostí v databázích, jehož cílem je extrakce vzorů dat. Metody lze rozlišit na průzkumné (exploratorní) a častěji používané prediktivní, které se pak označují jako prediktivní data mining. Součástí data miningu je vždy také příprava dat, která obsahuje jejich čištění, ošetření chybějících dat či transformace některých proměnných. Data miningové metody se často používají nejen v marketingu, ale také v molekulární genetice a medicíně. Mezi výhody data miningových metod patří schopnost vhodně kombinovat analýzu proměnných různého typu (spojité nebo kategoriální veličiny, případně i časové řady nebo obrazová informace), spolehlivost v situacích s větším procentem chybějících pozorování i spolehlivost při vysokých dimenzích dat, kdy klasické statistické metody selhávají.

Tento článek má za cíl prezentovat data mining jako užitečný přístup, který je obvyklý v medicínských aplikacích. I když nevyrůstá z klasických statistických předpokladů, v některých aplikacích nabízí spolehlivé výsledky, které by jen obtížně získaly jinými metodami. Kapitola 2 pojednává o aplikacích data miningových metod v medicínském rozhodování. Kapitola 3 se věnuje analýze asociací, která umožňuje provést extrakci vzorů pro kategoriální data. Kapitola 4 rozebírá vztah mezi neuronovými sítěmi a logistickou regresí, které jsou nejčastěji používanými data miningovými metodami [3]. Zároveň (oproti metodě z kapitoly 3) jde o běžně používané metody v informatice a statistice.

2. Systémy pro podporu klinického rozhodování

Tato kapitola má za cíl vysvětlit, že metody data miningu mají své důležité místo v medicínském výzkumu a při klinickém rozhodování. Nedílnou součástí klinického rozhodování je nejistota, která má spolu s daty a znalostmi také vliv na určení výsledné diagnózy, terapie i prognózy. Při procesu klinického rozhodování mohou asistovat systémy pro podporu rozhodování, což jsou velmi složité systémy schopné řešit široké spektrum různých úkolů, zpracovat informace různého typu a získat z nich konkrétní závěry. Porovnávají různé alternativy na základě porovnání jejich rizika a představují nedílnou součást technologií elektronického zdravotnictví (*e-health*).

Popíšeme konkrétní příklady použití data miningových metod pro aplikace v klinickém rozhodování.

Při studii rizikových faktorů aterosklerózy v české populaci mužů středního věku (`http://euromise.vse.cz/stulong-en/`) byla použita metoda GUHA (general unary hypotheses automaton), která byla navržena v článku [5] jako metoda pro automatické generování hypotéz na základě pozorovaných dat uložených v databázích. GUHA popisuje hypotézy pomocí vztahů mezi vlastnostmi objektů.

Práce [8] popisuje klinické systémy pro podporu rozhodování, jejich základní principy a strukturu a zaměřuje se na přínos takových systémů pro oblast bezpečnosti pacientů. Detailně popisuje principy moderních metod mnohorozměrné statistiky, které se používají pro klasifikační analýzu vysoce rozměrných dat v molekulární genetice.

V článku [2] jsou použity data miningové metody u systému pro podporu rozhodování pro praktické dětské lékaře. Článek vyšetřuje vztahy mezi důležitými klinickými veličinami, které souvisí s růstem dětí. V článku [11] jsou použity data miningové metody pro konstrukci systému pro podporu rozhodování, který umožňuje predikovat kardiovaskulární riziko u pacientů se selháním ledvin, kterým je dlouhodobě prováděna hemodialýza. Práce kombinuje shlukovou analýzu a metody založené na pravidlech (*rule based methods*), které jsou zaměřeny na porozumění struktuře dat a odvozování závěrů na základě logických pravidel. Článek [13] použil data miningové metody pro shlukovou analýzu pacientů, kteří trpí nedostatečnou funkcí štítné žlázy.

V poslední době se požaduje od systémů pro podporu klinického rozhodování, aby měly schopnost zpracovat informace různého typu. V článku [7] je popsán takový systém pro podporu rozhodování aplikovatelný na jednotce intenzivní péče, který zpracovává klinická data ve formě spojitých a kategoriálních proměnných, provádí analýzu časových řad a využívá i databázi znalostí zkonstruovanou experty.

Práce [6] popisuje text mining a jeho použití v systémech pro podporu klinického rozhodování. Jde o metodologii pro extrakci informace z textových dokumentů, vědeckých publikací či elektronického zdravotního záznamu jednotlivých pacientů, ale také klasifikaci dokumentů či shlukovou analýzu aplikovanou na databázi dokumentů. Za vstupní data je tedy považován text. Metody umožňují ze zadaného textu automaticky získat informace, mezi něž patří jména léků, názvy onemocnění, proteinů a genů, nebo provést přiřazení do již existujících medicínských ontologií. Často se text mining považuje za součást obecnějšího pojmu data mining, některé metody text miningu pocházejí z odlišných oborů, například z počítačové lingvistiky.

V následujících kapitolách se budeme podrobně zabývat některými z běžně používaných data miningových metod. Obecně je však třeba říci, že systémy pro podporu klinického rozhodování mohou najít široké uplatnění v rutinní lékařské péči teprve tehdy, až lékaři a další zdravotničtí pracovníci dosáhnou potřebné úrovně počítačové gramotnosti. Jejich vzdělávání vyžaduje výuku základů informačních věd a analytického využování spolu s teoretickými principy analýzy dat a rozhodování.

3. Extrakce vzorů kategoriálních dat

Algoritmus pro extrakci vzorů navržený v článku [15] představuje data miningovou metodu běžně používanou v medicínských aplikacích a je použit i v některých referencích citovaných v předchozí kapitole. Umožňuje analýzu složitých kvantitativních a kvalitativních asociací mezi proměnnými či jevy. Je použit na zpracování genetických dat v článku [14]. V této kapitole algoritmus popíšeme a převedeme jeho myšlenky do statistického jazyka.

Předpokládají se buď kategoriální data, anebo se spojité veličiny rozdělí tak, aby vznikla kategoriální veličina s malým počtem skupin. Taková kategoriální data tvoří kontingenční tabulku. Cílem je najít takové veličiny, které vytvářejí určitý vzorec (*pattern*), jinými slovy najít takové veličiny, které se signifikantně liší v různých situacích.

Metodu popíšeme na následující modelové studii, jejímž cílem je zjistit, které geny souvisejí se vznikem a rozvojem autoimunitního onemocnění štítné žlázy. Předpokládáme, že je k dispozici náhodný výběr pacientů s tímto onemocněním a na něm nezávislý náhodný výběr zdravých (kontrolních) osob, které netrpí žádným onemocněním štítné žlázy ani jiným autoimunitním onemocněním. U všech pacientů i kontrolních osob se provede odběr vzorku krve a pomocí technologie *microarrays* dojde k naměření genových expresí. Jde o hodnoty spojité veličiny, které odpovídají intenzitě aktivity jednotlivých genů v okamžiku odběru krve.

Uvažujme trojrozměrnou kontingenční tabulku $2 \times 2 \times 2$, která odpovídá počtům pacientů a zdravých osob v závislosti na hodnotách expresí dvou genů A a B. Vysoké exprese (větší než určitá konstanta δ) genů A a B budeme označovat pomocí $A = 1$ a $B = 1$, zatímco exprese menší než δ budou označeny jako $A = 0$ a $B = 0$. Tabulku četností nyní zapíšeme ve tvaru

	Nemocní		Zdraví	
	B=1	B=0	B=1	B=0
$A = 1$	n_{111}	n_{121}	n_{11K}	n_{12K}
$A = 0$	n_{211}	n_{221}	n_{21K}	n_{22K}

. (3)

Jistě by šlo data modelovat pomocí binární odezvy, která pro konkrétního pacienta vyjadřuje, zda je nemocný nebo zdravý. Budeme však uvažovat multinomický model. To znamená, že jednotlivým políčkům tabulky četností přísluší tabulka pravděpodobností

	Nemocní		Zdraví	
	$B{=}1$	$B{=}0$	$B{=}1$	$B{=}0$
$A=1$	π_{111}	π_{121}	π_{112}	π_{122}
$A=0$	π_{211}	π_{221}	π_{212}	π_{222}

, (4)

kde součet všech pravděpodobností je roven 1. Uvažujme nulovou hypotézu

$$H_0 : \pi_{ijk} = \pi_{i..}\pi_{.j.}\pi_{..k} \quad \text{pro konkrétní hodnoty } i,j,k \in \{1,2\}, \tag{5}$$

kde $\pi_{i..}$, $\pi_{.j.}$ a $\pi_{..k}$ jsou marginální pravděpodobnosti. Jde tedy o zformulování nulové hypotézy nezávislosti pouze pro jediné políčko dané tabulky s pevnými hodnotami i,j,k. Považujme n_{ijk} za realizaci náhodné veličiny s binomickým rozdělením $\mathsf{Bi}(n,\pi_{ijk})$, kde n označuje součet všech četností z tabulky (3).

Článek [15] uvažuje testovou statistiku založenou na normalizovaných reziduích (*adjusted residuals*) [1], která vzniknou normalizací (Pearsonových) reziduí na jednotkový rozptyl. Následně se využije jejich asymptotická normalita za platnosti H_0. Snadno se ověří, že tato metoda je ekvivalentní s následujícím postupem. Test H_0 proti oboustranné alternativě se provede na základě asymptotického vztahu

$$\frac{n_{ijk} - n\hat{\pi}_{ijk}}{\sqrt{n\hat{\pi}_{ijk}(1-\hat{\pi}_{ijk})}} \xrightarrow{\mathcal{D}} \mathsf{N}(0,1) \tag{6}$$

za platnosti H_0, kde $\hat{\pi}_{ijk}$ označuje maximálně věrohodný odhad pravděpodobnosti π_{ijk}. Ten za platnosti H_0 spočítáme jako

$$\hat{\pi}_{ijk} = \frac{n_{i..}n_{.j.}n_{..k}}{n^3}. \tag{7}$$

Alternativním postupem k asymptotickému testu by bylo využití zobecnění Fisherova faktoriálového testu pro vysoce rozměrná genetická data [10].

Článek [15] doporučil uvažovat testovou statistiku ze vzorce (6) pro různé hodnoty indexů $i,j,k \in \{1,2\}$. Přitom je však jistě žádoucí uvažovat i testovou statistiku například pro tabulku

	Nemocní	Zdraví
$A=1$	$n_{1\cdot1}$	$n_{1\cdot2}$
$A=0$	$n_{2\cdot1}$	$n_{2\cdot2}$

, (8)

která vznikne z tabulky (3) tak, že zcela ignorujeme efekt genu B. Proto se v praxi postupuje tak, že se uvažují také různé tabulky, které vzniknou z původní tabulky (3) jako marginální tabulky ignorováním vlivu některé z kategoriálních proměnných. Tímto způsobem metoda odhalí i jednodušší asociace mezi proměnnými (asociace nižších řádů). Takový postup však nebere v úvahu, že se jedná o mnohonásobné testování. Z toho plyne, že celý postup extrakce vzorů nedrží pravděpodobnost chyby 1. druhu.

4. Klasifikační metody

Řada z metod, které bývají používány při data miningu, má za cíl sestavit klasifikační pravidlo pro mnohorozměrná data tak, aby bylo možno automaticky zařazovat nová pozorování do dvou nebo více skupin. Jedná se o běžně používané klasifikační metody, kterým se však v informatickém kontextu říká metody strojového učení (*machine learning*). Zde se zastavíme u logistické klasifikace a popíšeme jeden speciální případ neuronové sítě, který je přesně roven modelu logistické regrese. Obecně lze tvrdit, že neuronové sítě představují přirozené zobecnění logistické regrese [3].

4.1. Logistická klasifikace

Logistická regrese je metodou pro regresní modelování binární odezvy. Označíme pomocí $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ binární odezvu, jejíž hodnoty 1 (resp. 0) se interpretují jako zdar (resp. nezdar), tedy situaci, kdy nastává (resp. nenastává) nějaký uvažovaný jev. Pravděpodobnost zdaru pro i-té pozorování se modeluje jako odezva nezávisle proměnných $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip}$, které mohou být spojité i diskrétní.

Označme pomocí π_i pravděpodobnost zdaru pro i-té pozorování. Uvažujeme model logistické regrese s absoutním členem, tedy

$$Y_i \sim \mathsf{Alt}(\pi_i), \quad \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \tag{9}$$

kde $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ tvoří vektor regresních parametrů.

Popíšeme logistickou klasifikační analýzu do dvou skupin. Uvažujeme dva nezávislé náhodné výběry p-rozměrných dat. Definujeme odezvu $\mathbf{Y}$ jako binární proměnnou, která pro i-té pozorování vyjadřuje indikátor jevu, zda pozorování $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ pochází či nepochází z první skupiny. V modelu (9) se odhadnou regresní parametry pomocí $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$. Pomocí

π^* označme

$$\pi^* = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 Z_1 + \cdots + \hat{\beta}_p Z_p\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 Z_1 + \cdots + \hat{\beta}_p Z_p\}}. \tag{10}$$

Ve statistických učebnicích se obvykle doporučuje klasifikovat nové pozorování $\mathbf{Z}$ do 1. skupiny tehdy a jen tehdy, když $\pi^* > 1/2$; viz například [12]. Přitom ale v některých případech může dojít k velmi nevýhodné situaci, kdy se jedné z obou skupin klasifikuje drtivá většina trénovacích dat. K tomu dochází při velmi odlišných počtech pozorování v obou skupinách. Proto je žádoucí nahradit klasifikační pravidlo $\pi^* > 1/2$ pravidlem $\pi^* > c$, kde konstanta c je určena tak, aby byla minimální klasifikační chyba. Výhodnou možností je maximalizovat Youdenův index Y definovaný pomocí

$$Y = \text{senzitivita} + \text{specificita} - 1. \tag{11}$$

Zkušenost ukazuje, že taková optimální konstanta c může být výrazně odlišná od $1/2$.

4.2. Neuronové sítě

Neuronové sítě (*neural networks*) byly navrženy jako modely pro nervové buňky (pro biologické aplikace). Zatímco logistická regrese uvažuje modely s malou komplexitou, u neuronových sítí často nastává přeučení (*overfitting*), protože mohou obsahovat velké množství parametrů. U neuronových sítí ani nelze testovat významnost parametrů. Je typické, že se neuronové sítě označují jako černé skříňky s velkým množstvím parametrů, které nelze jednoznačně interpretovat. Pro nalezení jejich vhodných odhadů se vyžaduje hodně velký počet pozorování. Naproti tomu u logistické regrese se hovoří spíše o bílé skříňce (*white-box model*), protože nabízí jednodušší interpretaci. Proto se v medicínském výzkumu logistická regrese používá častěji než neuronové sítě, které mohou být považovány za neparametrickou (a tedy výrazně složitější) metodu.

Existují různé druhy neuronových sítí, které vykazují velkou flexibilitu. Zde se zabýváme supervizovanými sítěmi pro klasifikaci do dvou skupin. Supervizované klasifikační metody při procesu učení (při formulaci klasifikačního pravidla) využijí informaci o tom, do které skupiny patří jednotlivá pozorování z trénovací množiny dat. Cílem supervizovaných metod tedy je popsat (modelovat) odlišnost mezi jednotlivými (pevně danými) dvěma skupinami.

Neuronová síť se skládá ze vstupní a výstupní vrstvy neuronů, případně jedné nebo více skrytých vrstev, jež jsou navzájem propojeny pomocí hran.

Váhy propojující každý neuron s některými z neuronů z další vrstvy se určují v průběhu procesu učení. Neuronová síť má za svůj výstup hodnotu aktivační (přenosové) funkce spočítanou pro vážené vstupy.

Uvažujeme dva nezávislé náhodné výběry p-rozměrných dat. Jednotlivým vstupům (proměnným) přísluší takzvané váhy $\mathbf{w}$, které mohou nabývat libovolných reálných (i záporných) hodnot; jde vlastně o běžné regresní parametry. Výstupem sítě pak je hodnota $f = g(\mathbf{w}^T\mathbf{x} + b)$, kde b je konstanta (absolutní člen). Podle charakteru výstupu lze rozlišit neuronové sítě spojité a binární. U jednoduchých neuronových sítí se za funkci g nejčastěji volí ryze monotónní funkce, mezi něž patří logistická funkce nebo hyperbolický tangens.

Neuronová síť bez vnitřních vrstev s logistickou aktivační funkcí

$$g^*(x) = \frac{1}{1 + e^{-x}}, \quad x \in \mathbb{R}, \tag{12}$$

je přesně rovna modelu logistické regrese. Tento speciální případ neuronové sítě se od logistické regrese odlišuje pouze metodou pro odhad (regresních) parametrů.

Hyperbolický tangens souvisí s logistickou funkcí pomocí vztahu

$$\tanh(x) = 2g^*(2x) + 1, \quad x \in \mathbb{R}. \tag{13}$$

Snadno se ověří, že klasifikace založená na zobecněném lineárním modelu se spojovací funkcí (13) je ekvivalentní logistické klasifikaci. Odtud plyne závěr, že neuronová síť bez vnitřních vrstev s aktivační funkcí (13) dává identické klasifikační výsledky jako táž síť s logistickou aktivační funkcí. Pouze u sítí s jednou či více vnitřními vrstvami není ekvivalence mezi aktivační funkcí (12) a (13).

Popišme ještě metody pro odhad parametrů u neuronových sítí a srovnejme ji s odhadováním parametrů logistické regrese. Metoda zpětné propagace (*back-propagation*) je obvyklou metodou pro odhad parametrů v neuronové síti, i když existují i méně běžné neuronové sítě, které využívají maximální věrohodnosti. Zpětná propagace požaduje, aby byly určeny počáteční odhady parametrů, což jsou váhy jednotlivých uzlů neuronové sítě. Při dopředném průchodu sítí se postupně počítají váhy neuronů v dalších vrstvách, až je možné spočítat hodnota výstupu a odtud i celkovou klasifikační chybu přes celou trénovací množinu dat. V další iteraci je snahou tuto klasifikační chybu zmenšit. Proto metoda prochází celou sítě zpětně, přičemž na základě hodnoty chyby se upraví váhy pro jednotlivé uzly sítě. Přitom se používá optimalizační metoda největšího spádu. Celkově se tedy iterativně odečítá

určitý násobek gradientu vah od počátečních vah. Jiná je situace u logistické regrese, kde se parametry odhadují pomocí metody maximální věrohodnosti. Ta se ovšem převede na úlohu hledání kořene složité nelineární funkce, pro niž se aplikuje Newtonova(-Raphsonova) metoda.

5. Závěr

V četných referencích byly popsány různé systémy pro podporu klinického rozhodování spolu s jejich výsledky, které jsou dosaženy na reálných medicínských datech. Některé systémy založené na data miningových metodách jsou připravené, aby mohly plnit svou asistenční roli při rutinní lékařské péči, kdy mohou pomoci při stanovení diagnózy, terapie i prognózy u jednotlivých pacientů. S použitím data miningových metod byly například v poslední době v medicíně odhaleny zákonitosti pro genetickou podmíněnost některých běžných onemocnění. Očekává se, že budou brzy odvozeny obdobné výsledky pro celou řadu dalších onemocnění. Výsledky data miningových metod se v odborné literatuře považují za důvěryhodné i přesto, že mohou být kritizovány ze statistického hlediska za nesplnění předpokladů. Zároveň je však třeba přiznat, že stejnou výtku lze vznést i vůči celé řadě dalších informatických metod, které jsou běžně používány; příkladem mohou být heuristické postupy používané při analýze obrazové informace.

Přestože data miningové metody bývají zařazovány do informatiky, lze říci, že souvisí se statistikou či přímo statistiku využívají. Domnívám se, že z toho důvodu by se měla ve statistické komunitě věnovat data miningovým metodám větší pozornost.

Literatura

[1] Agresti A. (2002): *Categorical data analysis.* Second edition. Wiley, New York.

[2] Downs S. M., Wallace M. Y. (2000): Mining association rules from a pediatric primary care decision support system. *Proceedings American Medical Informatics Association Symposium 2000*, 200 – 204.

[3] Dreiseitl S., Ohno-Machado L. (2002): Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics* **35**, 352 – 359.

[4] Gorunescu F. (2011): *Data mining: concepts, models and techniques.* Springer, Berlin.

[5] Hájek P., Havel I., Chytil M. (1966): GUHA – metoda systematického vyhledávání hypotéz. *Kybernetika* **2** (1), 31 – 47.

[6] Chen H., Fuller S. S., Friedman C., Hersh W. (2005): *Medical informatics. Knowledge management and data mining in biomedicine.* Springer, New York.

[7] Imhoff M., Gather U., Morik K. (1999): Development of decision support algorithms for intensive care medicine: A new approach combining time series analysis and a knowledge base system with learning and revision capabilities. In Burgard W., Christaller T., Cremers A. B. (Eds.): KI-99, Advances in Artificial Intelligence, *Lecture Notes in Artifical Intelligence* 1701, Springer, Berlin, 219 – 230.

[8] Kalina J., Zvárová J. (2012): Decision support systems in the process of improving patient safety. In Moumtzoglou A., Kastania A. (Eds.): *E-Health technologies and improving patient safety: Exploring organizational factors.* IGI Global, Hershey, Pennsylvania. Zasláno.

[9] Klímek J. (2005): Úvod do problematiky data miningu. *Informační bulletin České statistické společnosti* **16** (3), 12 – 19.

[10] Malaspinas A.-S., Uhler C. (2011): Detecting epistasis via Markov bases. *Journal of algebraic statistics* **2** (1), 36 – 53.

[11] Pfaff M., Weller K., Woetzel D., Guthke R., Schroeder K., Stein G., Pohlmeier R., Vienken J. (2004): Prediction of cardiovascular risk in hemodialysis patients by data mining. *Methods of Information in Medicine* **43** (1), 106 – 113.

[12] Stankovičová I., Vojtková M. (2007): *Viacrozmerné štatistické metódy s aplikáciami.* Iura edition, Bratislava.

[13] Temurtas F. (2009): A comparative study on thyroid disease diagnosis using neural networks. *Expert Systems with Applications* **36** (1), 949 – 949.

[14] Wong A. K. C., Au W.-H., Chan K. C. C. (2008): Discovering high-order patterns of gene expression levels. *Journal of Computational Biology* **15** (6), 625 – 637.

[15] Wong A. K. C., Wang Y. (1997): High-order pattern discovery from discrete-valued data. *IEEE Transactions on knowledge and data engineering* **9** (6), 877 – 893.

[16] Žváček J. (2007): Statistické výpočetní prostředí 2007. *Informační bulletin České statistické společnosti* **18** (3), 1 – 15.

NEUVĚŘITELNÉ ŠTĚSTÍ A NEUVĚŘITELNÁ SMŮLA BÝVAJÍ NĚKDY DOCELA UVĚŘITELNÉ

Ondřej Vencálek

Adresa: ÚPOL, KMA, 17. listopadu 1192/12, 771 46 Olomouc

E-mail: ondrej.vencalek@upol.cz

Abstrakt

Řídké jevy vždy fascinovaly, a nejenom davy. Mnohdy přitom byly důvodem k nečekaným objevům. A i o tom je mimo jiné tato úvaha.

Anomalous events often lie in the roots of discoveries in science and of actions in other domains. Familiar examples are discovery of pulsars, the identification of initial signs of an epidemic, and the detection of faults and fraud. In general, they are events which are seen as so unexpected or improbable that one is led to suspect there must be underlying causes. However, to determine whether such events are genuinely improbable, one needs to evaluate their probabilities under normal conditions. It is all too easy to underestimate such probabilities.

1. Neuvěřitelné štěstí

Čas od času se z médií dozvídáme opravdu senzační zprávy. Před nedávnem to byla zpráva o neuvěřitelném štěstí jednoho golfisty. Zněla takto: „Britský penzista Peter Wafford (75) se vytáhl na golfovém turnaji seniorů v Chigwellu v hrabství Essex. Dvě tříparové jamky po sobě totiž trefil hned na první pokus. Pravděpodobnost něčeho takového je asi 67 miliónů ku jedné. To je asi tak mizivá šance, jako že vyhrajete hlavní cenu v loterii několikrát po sobě.“

Fascinující! Zajisté, jenže... Je opravdu toto štěstí tak neuvěřitelné? Připusťme, že uvedená pravděpodobnost je skutečnou pravděpodobností sledovaného jevu. Mimochodem, u takového čísla člověk zřejmě vždy lehce zaváhá, jakým způsobem bylo získáno, tedy kde se vlastně vzalo. Zde zřejmě někdo z dlouhodobých statistik zjistil, že „tříparová jamka“ je na první pokus trefena přibližně v jednom z asi 8200 pokusů a umocněním pak došel k odhadu, že pravděpodobnost „neuvěřitelného štěstí“, jaké měl Peter Wafford, je asi jedna ku 67 miliónům. Nechť tedy toto číslo vyjadřuje skutečnou pravděpodobnost sledovaného jevu. Je malinká, to ano, ale při úvahách o výskytu sledovaného jevu musíme vždy zároveň uvažovat počet pokusů, které jsme učinili. Pokud jste, stejně jako já, až dosud o golfu nevěděli vůbec nic, pak

vězte, že podle České golfové federace „golf patří mezi tři nejčastěji provozované sporty na světě. Aktivně se mu věnuje více než 70 miliónů lidí ve více než 120 zemích.“ Pokud bychom za aktivní účast hráče považovali alespoň jednu dvojici odpalů na tříparovou jamku za rok, došli bychom k závěru, že pravděpodobnost toho, že alespoň jeden hráč bude mít během sledovaného roku „neuvěřitelné štěstí“ je přibližně 0,65. Je tedy poměrně velká a občasný výskyt „neuvěřitelného štěstí“ není zase až tak překvapivý.

Ostatně nikoho z nás nepřekvapí, když čas od času někdo vyhraje první cenu ve sportce, i když každý, kdo alespoň někdy viděl Andělovu knihu *Matematika náhody,* ví, že pravděpodobnost výhry první ceny ve sportce je přibližně jedna ku 14 miliónům. Jedná se tedy řádově o stejné štěstí jako ve výše uvedeném golfovém případě.

2. Podivné náhody

Kromě neuvěřitelného štěstí jsou i jevy, které bychom štěstím přímo nenazvali, ale které nám přijdou málo pravděpodobné, ba téměř nemožné, a jejichž výskyt považujeme za cosi podivného, někdy i podezřelého.

Ve výše uvedené knize *Matematika náhody* se uvádí případ, kdy ve volbách do Poslanecké sněmovny Parlamentu ČR roku 1998 získaly dvě politické strany naprosto stejný počet hlasů ve třech různých okresech (1295 v okrese Domažlice, 2664 v okrese Karlovy Vary a 2105 v okrese Uherské Hradiště). Za docela realistických předpokladů přitom Anděl dospěl k odhadu pravděpodobnosti stejného počtu hlasů v jednom okrese p=0,0081. A pravděpodobnost, že sledované strany získají stejný počet hlasů alespoň ve třech volebních okresech pak vyčíslil na 0,036, což je „pravděpodobnost malá, ale ne zcela zanedbatelná“. Jinými slovy opět se nejedná o jev, který by byl „neuvěřitelný“. Poznamenejme jen, že skeptika, který ví, že o náhodu v našich zeměpisných šířkách rozhodně jít nemohlo, výpočet nepřesvědčí.

Zdá se, že profesor Anděl si v podobných příkladech libuje. V Bulletinu z roku 2005 publikoval spolu s docentem Zvárou článek nazvaný *Náhodné tipy ve sportce.* V něm vyčíslují pravděpodobnost toho, že při náhodné volbě šesti čísel z množiny $\{1, \ldots, 49\}$ alespoň dvě z nich budou sousedit. Docházejí k číslu 0,495. Poznamenejme jen, že skeptika, který ví, že generátor čísel společnosti hru provozující rozhodně není náhodný, ani tentokrát výpočet nepřesvědčí.

Mimochodem víte, jaká je pravděpodobnost, že dva Pražáci mají naprosto stejný počet vlasů? Důsledný otázaný začne uvažovat o rozdílnostech mezi miminky, dospělými a starci nebo mezi muži a ženami. Přitom ho jistě napadne, že někteří muži mají pleš, někteří jsou zcela holohlaví (hle, jasná

odpověď se nabízí). Ti méně důslední, kteří si představí populaci Pražanů jako zcela homogenní, většinou ví, že naše hlavní město má asi jeden a čtvrt miliónu obyvatel, a tak jim k vytvoření hrubé představy o hledané pravděpodobnosti stačí doplnit informaci, kolik tak obvykle člověk mívá vlasů. Z různých zdrojů se dozvídáme, že je to přibližně sto tisíc. (Uvádí se mezi 80 a 140 tisíci). Nemůžeme pak dojít k jinému závěru, než že správná odpověď na výše uvedenou otázku je p=1,00. Jistě je totiž rozumný předpoklad, že aspoň milión obyvatel nemá více než 200 000 vlasů (což je dvojnásobek oproti normálu). Vytvoříme-li si tedy tabulku o 200 000 řádcích a do každého z nich napíšeme jména těch z miliónu obyvatel, kteří mají právě tolik vlasů, jako je pořadové číslo tohoto řádku, pak jistě v alespoň jednom řádku bude více než jedno jméno.

3. Neuvěřitelná smůla (zákon schválnosti)

U nás doma se příležitostně hrávala hra Šťastných deset. V ní je z osmdesáti čísel taženo dvacet. Sázející přitom může sázet (maximálně) deset čísel. Krom občasných výher v řádu desítek či stovek korun jsme občas mívali neuvěřitelnou smůlu. Kdybychom totiž místo sedmičky vsadili na osmičku, místo 15 na 14, místo 58 na 59 a místo 63 na 62, byli bychom bývali vyhráli první cenu! Byli jsme tak blízko! Zákon schválnosti však zapůsobil vždy v náš neprospěch.

Jak si tu smůlu vysvětlit? Budeme-li počítat mimo deseti námi vsazených čísel také čísla o jedničku větší a menší, budeme mít „vsazeno“ nikoliv deset, ale mnohem více (až třicet) čísel. Jednoduchá simulace nám ukáže, že těchto čísel bude nejčastěji mezi 24 až 27. Není pak těžké rozšířit simulaci tak, abychom získali odhad pravděpodobnosti, že bude taženo alespoň deset čísel z těch, která jsme vsadili a čísel o jedna větších či menších. Tuto pravděpodobnost můžeme na základě simulace odhadnout číslem p=0,051. Jistě nás pak nepřekvapí, že při jednom tahu každý den jsme měli minimálně jednou do měsíce neuvěřitelnou smůlu. Pravděpodobnost toho, že tato smůla nastane v aspoň jednom ze třiceti nezávislých tahů je přibližně $1-(1-0{,}051)^{30} = 0{,}79$.

Mimochodem všimněte si, že jsme mohli „blízkými čísly“ uhodnout všech dvacet tažených čísel, aniž bychom skutečně vsazenými čísly jedinkrát správně uhodli. To už by byla opravdu super-neuvěřitelná smůla.

4. Čím to je

Všechny výše uvedené příklady mají společné to, že se zabývají jevy „neuvěřitelnými“, tedy jevy s velmi malou pravděpodobností. Zároveň se však zabývají otázkami typu: jaká je pravděpodobnost, že sledovaný jev nastane

alespoň jednou. Přitom však možností, kdy sledovaný jev může nastat, bývá hodně. Všem „jasný“ závěr, že při hodně pokusech se i jev s malou pravděpodobností tu a tam vyskytne, je pro nás v běžném životě často obtížně akceptovatelný. Některé „náhody“ nám přijdou podezřelé. Toto podezření nás vede k používání termínů jako zákon schválnosti či dokonce spiknutí. A přitom vysvětlení může být docela prosté a racionální.

Ještě o jednom jevu stojí za to se zmínit. Podezřelé souvislosti totiž člověk nachází vždy až zpětně. Teprve když ve sportce vyjdou čísla 1, 2, 3, 4, 5, 6, začneme se bouřit proti náhodnosti losování. Přitom víme, že tato kombinace není při náhodném tahu o nic víc, ale také o nic míň pravděpodobná, než kterákoliv kombinace jiná, například 7, 9, 16, 26, 40, 47 (šestice tažená v prvním tahu dne 26. 5. 2010, nijak podezřelá). Toto však neplatí jen pro matematiku. Kolikrát slýcháme: „Já si celý den říkal, že se zrovna něco důležitého stane, a vidíte, stalo se . . . “. Ten člověk má jistě neuvěřitelnou intuici! Vždyť kolik je dnů, kdy se opravdu něco stane?

PROFESOR LUBOMÍR KUBÁČEK OSEMDESIATNIK

Marie Hušková a Júlia Volaufová

Adresa: KPMS MFF UK, Sokolovská 83, 186 00, Praha 8
LSU Health Sciences Center, School of Public Health, 2020 Gravier Street, New Orleans, LA 70112

E-mail: huskova@karlin.mff.cuni.cz, jvolau@lsuhsc.edu

Abstrakt

Článek je věnován významnému životnímu jubileu profesora Lubomíra Kubáčka.

Professor Lubomír Kubáček has extensively contributed to many areas of statistics. His contribution is highlighted here with reverend appreciation and admiration from his colleagues.

Je neuveriteľné, že náš priateľ a kolega, profesor Luboš Kubáček, sa už dožíva osemdesiatky. Dožíva sa jej v plnom zdraví, plný vitality a pracovnej energie.

Narodil sa 1. februára 1931 v Bratislave. Po absolvovaní štúdia geodézie na Slovenskej vysokej škole technickej v r. 1954 nastúpil ako vedúci výpočtového oddelenia v Geodetickom ústave v Bratislave. Zotrval tam osem rokov.

Prostredie čísel a presných výpočtov ho silne motivovali a utvrdili v presvedčení, že bez matematiky – numerických metód a štatistiky – to ďalej nepôjde. Začal teda študovať popri zamestnaní a úspešne ukončil najprv v r. 1957 matematickú analýzu a v r. 1964 pravdepodobnosť a matematickú štatistiku, obe na Prírodovedeckej fakulte Univerzity Komenského v Bratislave. Neskôr prešiel do Slovenskej Akadémie Vied, do ústavu teórie merania. Tu sa profesorovi Kubáčkovi podarilo, vďaka jeho obrovskému elánu, vybudovať silný kolektív štatistikov a matematikov, súčasť tzv. oddelenia teoretických metód. Z tohto kolektívu vyšli mnohí slovenskí štatistici a matematici, pôsobiaci na slovenských ale aj zahraničných akademických pracoviskách. Pravidelné stretnutia na seminároch ako magnet priťahovali mladých ľudí z celého Slovenska. Niektorí neváhali a pravidelne cestovali aj zo vzdialenejších miest.

V r. 1981 profesor Kubáček prešiel do Matematického ústavu SAV, kde v r. 1988–91 bol jeho riaditeľom, ale ani v období riaditeľovania nepoľavil v úzkej spolupráci s kolegami a mladými študentami.

Počas celého pôsobenia v Bratislave zostal verný geodézii. Jeho teoretické práce, knižné publikácie a vedecké články, v ktorých sa venuje najmä riešeniu obtiažnych problémov v oblasti regresných modelov, a vďaka ktorým získal medzinárodné uznania, majú vždy priamu nadväznosť na konkrétne aplikácie. Napríklad veľmi výrazne prispel k rozvoju štatistiky teórie geodetických sietí. Neskôr sa však jeho aplikačný obzor ešte viac rozšíril – počas zhruba dvadsiatich rokov spolupracoval na riešení medicínskych a biomedicínskych problémov s 1. Internou klinikou v Bratislave. Spolu s manželkou Liduškou, ktorá neochvejne stála po jeho boku a bola mu životnou partnerkou a najbližšou spolupracovníčkou až do jej smrti, prispeli k riešeniu mnohých teoretických štatistických problémov v geofyzike. V tom období, v r. 1981 získal titul DrSc. a v r. 1991 bol menovaný profesorom.

Od roku 1994 profesor Kubáček pôsobí – pracuje, publikuje, prednáša a venuje sa naďalej veľmi aktívne výchove mladých matematikov – na Prírodovedeckej fakulte Univerzity Palackého v Olomouci. Aj tu sa jeho charizma naplno prejavila. Veľkou mierou sa zaslúžil o vybudovanie štatistickej skupiny v rámci aplikovanej matematiky.

Už počas pôsobenia v ÚTM SAV v Bratislave, profesor Kubáček pravidelne prednášal pravdepodobnosť a matematickú štatistiku na Komenského univerzite v Bratislave a venoval sa výchove ašpirantov a mladých vedeckých pracovníkov. Jeho obetavosť a ochota pomáhať mladým nepozná hranice. Dodnes, hoci po X-tý krát dokáže tráviť mnohé hodiny s mladým adeptom alebo adeptkou a zasväcovať ich trpezlivo krok po kroku do základov pravdepodobnosti a štatistiky. Podarilo sa mu vyškoliť, či už na Slovensku alebo

v Čechách najmenej 15 doktorandov, pričom v súčasnosti školí ďalších dvoch nádejných adeptov matematických vied.

Profesor Kubáček dodnes nepoľavil v základnom teoretickom výskume – výsledky publikoval a stále publikuje nielen vo vedeckých časopisoch, ale v celom rade vedeckých kníh. Je autorom alebo spoluautorom 11 odborných kníh, z toho dva boli publikované renomovaným zahraničným nakladateľstvom. Je autorom alebo spoluautorom 6 skrípt a viac než 130 článkov v uznávaných medzinárodných vedeckých časopisoch. Odborné publikácie nájdeme v matematicky zameraných časopisoch aj časopisoch zameraných na geodéziu, chemometriu a lekársky výskum. Je autorom celého radu popularizačných článkov. Je členom niekoľkých redakčných rád vedeckých a odborných časopisov. Je nositeľom celého radu medailí, vyznamenaní a ocenení za celoživotnú prácu, za rozvoj matematickej štatistiky, aplikácií a popularizácie matematiky na Slovensku, v Čechách a vo svete.

Výbor ČStS praje pevné zdravie, mnoho nových plodných myšlienok, nápadov a riešení, a najmä hodne spokojnosti v kruhu svojich kolegov, priateľov a najbližšej rodiny.

Nitrianske štatistické dni

Dagmar Markechová

E-mail: dmarkechova@ukf.sk

V dňoch 27. a 28. mája 2010 sa na Katedre matematiky Fakulty prírodných vied Univerzity Konštantína Filozofa v Nitre pod záštitou dekana FPV UKF v Nitre, prof. RNDr. Ľubomíra Zelenického, CSc., uskutočnila v poradí druhá medzinárodná konferencia Nitrianske štatistické dni. Konferenciu zorganizovala Katedra matematiky FPV UKF v Nitre v spolupráci so Slovenskou štatistickou a demografickou spoločnosťou. Dekan FPV UKF v Nitre konferenciu slávnostne otvoril a pri tejto príležitosti odovzdal Dr.h.c. prof. RNDr. Beloslavovi Riečanovi, DrSc. Pamätnú medailu FPV UKF v Nitre.

Tematicky bola konferencia venovaná aktuálnym trendom matematickej štatistiky, teórie pravdepodobnosti a analýzy dát, aplikáciám štatistiky a výučbe štatistiky. Účastníci konferencie mali skvelú možnosť stretnúť sa a vypočuť si prednášky vzácnych hostí, prof. B. Riečana a prof. J. Antocha. Cieľom prednášky prof. B. Riečana (Probability on algebraic structures) bolo vzbudiť záujem o výskum v oblasti, v ktorej bola práve vyvinutá nová metóda. Táto metóda bola aplikovaná na MV- algebry a čiastočne tiež na D- posety.

Prof. J. Antoch sa vo svojej veľmi zaujímavej prednáške zameral na spracovanie dát z oblasti životného prostredia a na hľadanie zmien v štatistických modeloch (Change point detection).

Príspevky, ktoré sme si mohli v priebehu konferencie vypočuť, boli veľmi rozmanité a boli z rôznych oblastí. Napríklad doc. J. Chajdiak tu vystúpil s príspevkom „Stupeň dôležitosti zdrojov informácií pri inovačných aktivitách“. Prof. M. Bauerová, doc. Brindza, prof. B. Stehlíková a prof. A. Tirpáková svoje dva príspevky venovali kvantifikácii biodiverzity a porovnávaniu mier biodiverzity s využitím štatistických metód. RNDr. J. Luha hovoril o analýze odpovedí „neviem“ v batérii otázok. RNDr. Ľ. Rybanský sa zaoberal modelmi výpočtu pravdepodobnosti z kurzu. Príspevky Ing. J. Juriovej, Ing. S. Kapounka, Ing. Ľ. Fabovej, R. Martinákovej, Ing. L. Muru, RNDr. J. Poměnkovej, doc. Ing. R. Maršálka, Mgr. M. Řezáča, doc. M. Urbaníkovej, Ing. Z. Polákovej a doc. P. Obtuloviča boli venované aplikáciám metód matematickej štatistiky v ekonómii resp. vo finančnom sektore. RNDr. O. Kříž sa zaoberal výučbou štatistiky podporovanou excelovskou aplikáciou. Doc. J. Broďáni hovoril o prognózovaní v športe.

Príspevky z konferencie sú publikované v 2. čísle šiesteho ročníka časopisu Forum Statisticum Slovacum.

Diskusia, ktorá sa rozhodne netýkala iba tematických okruhov konferencie, sa presunula vo večerných hodinách do neďalekej kaviarne.

Záverom možno povedať, že Nitrianske štatistické dni 2010 sa vydarili po všetkých stránkach. Zároveň by sme chceli vysloviť prianie, aby sa v tomto trende pokračovalo aj v ďalších ročníkoch Nitrianskych štatistických dní.

ISI Young Statisticians Meeting, Dublin 2011

Lukáš Pastorek

E-mail: lukas.pastorek@vse.cz

Ve dnech 19. – 21. srpna se při příležitosti konání celosvětového kongresu ISI 2011 v Irsku uskutečnilo satelitní setkání mladých statistiků na půdě více než 400 let staré Trinity College v centru Dublinu. Setkání, kterého se účastnili i mladí statistici z České a Slovenské republiky, mělo za úkol aktivně zapojit statistiky v raném stádiu jejich kariéry prostřednictvím posterových prezentací jejich dosavadní práce a účastí na přednáškách popředních statistiků.

Předsednictví této akce se ujal Victor M. Panaretos ze Švýcarska, jakožto nadějná vědecká hvězda na statistickém nebi a zároveň nejmladší volený člen

ISI. Účastníky také uvítal Jef Teugels z titulu prezidenta ISI, který připomněl důležitost a nevyhnutelnost podpory nových mladých statistiků. Celkově se podařilo vytvořit uvolněnou neformální atmosféru, která byla živnou půdou pro navazování nových profesionálních vztahů a prezentování svého vlastního výzkumu před zraky svých mladých kolegů.

V průběhu konání této akce mělo více než 150 účastníků z celého světa možnost zhlédnout prezentace ku příkladu profesora Adriana Raftery z Univerzity ve Washingtone s prezentací o pravděpodobnostních modelech při předpovídání počasí. Raftery zdůraznil mezery ostatních oborů, které mohou a dokážou statistikové svým přístupem zaplnit. Nebo Rajna Patela z Google Research, který se zaměřil na ilustraci problémů při optimalizaci jejich vyhledávacího algoritmu. Martin Wainwright z Berkley zase poukázal na praktických ukázkách na nástrahy vícerozměrných prostorů a metody jejich „podmanění“. Organizační výbor dal možnost i výhercům ceny Jana Tinbergena (M. Roozbeh z Iránu, Manjule Kalluraya z Indie a Kodzovi Senu Abalo z Pobřeží Slonoviny), která je udělovaná mladým statistikům z rozvojových zemí, odprezentovat výsledky své práce před svými mladými kolegy. Přednáškové pásmo uzavřel svou prezentací Sir David Cox na téma Souhry teorie a aplikace v statistice, čím otevřel v publiku diskuzi na „věčné“ téma Bayes vs. Frekvencionisté.

V posterových sekcích se objevili práce nejrůznějších teoretických nebo aplikačních zaměření. Účastníci přitom neváhali a využili tuto možnost ke komunikaci se svými mladými statistickými „druhy“ a obeznámili se s jejich výzkumem. Bohatá diskuze, která se rozvinula před přednáškovými sály v každé posterové sekci, překonala očekávání pořadatelů. Z České republiky odprezentoval svůj poster o neparametrických přístupech k detekci změn v rozdělení Ondrej Chochola z MFF UK a také Lukáš Pastorek a Tomáš Vintr z FIS VŠE, kteří se zaměřili na využití shlukování při navigaci autonomního robotu.

Konec setkání se nesl v duchu vyhlašování nejlepších posterů, za které výherci obdrželi kromě zaslouženého potlesku i finanční odměnu.

YSI Dublin nebylo jenom o pasivní anonymní účasti, jak to na mnohých konferencích bývá. Bylo to hlavně o zapojení se a aktivním sdílení výsledků své práce, za kterou si člověk stojí a které věnuje denně čas, s ostatními mladými souputníky. Navzdory „náročnosti“ navazování profesionálních vztahů na větších akcích, setkání tohoto typu dokáže překonat tyto pomyslné bariéry, které máme v mysli, a může nás postrčit směrem, který nám může otevřít úplně nové dimenze spolupráce.

NAKLADATELSTVÍ INFORMATORIUM

Olga Hebáková

Adresa: Informatorium, Mandova 449/14, 149 00 Praha 11

E-mail: hebakova@informatorium.cz

Abstrakt

Nakladatelství Informatorium vzniklo začátkem roku 1991 a je zaměřeno převážně na vydávání středoškolských učebnic, odborné a populárně naučné literatury.

Publishing house Informatorium has been established in 1991. It is oriented especially on the publishing secondary schools textbooks, professional and infotaintment literature.

Za dobu své činnosti Informatorium vydalo téměř 320 publikací, a to z oblasti archeologie, dřevařství, ekologie, ekonomie, elektrotechniky, chemie, jazykových učebnic, kadeřnictví, keramiky, kosmetiky, potravinářství, práva, psychologie, rybářství, sklářství, stavebnictví, strojírenství, textilu, zdravotnictví, zemědělství a dalších oborů. Převážný podíl z tohoto počtu tvoří středoškolské odborné učebnice.

V koedicích se zahraničními nakladateli vydalo rovněž řadu vysoce odborných monografií v angličtině, němčině, švédštině a francouzštině.

Dne 23. listopadu 2011 byla ustavena Vědecká redakce nakladatelství Informatorium a byli jmenováni její členové. Zkušenosti a znalosti členů vědecké redakce je zárukou splnění všech požadavků kladených na odborné knihy a zabezpečení kvality vědecky zaměřených publikací určených k vydání a distribuci na knižním trhu. Úkolem vědecké redakce je vybrat a jmenovat vhodné oponenty a na základě jejich lektorských posudků a vlastních zkušeností rozhodnout, zda text vyhovuje všem požadavkům. Ekonomická kritéria jsou pro vydávání knih pro každé nakladatelství vždy na prvním místě, ale z odborného hlediska bude rozhodnutí vědecké redakce plně respektováno.

Členové Vědecké redakce nakladatelství Informatorium, spol. s r.o.:

- prof. Ing. Václav Čermák, DrSc.
- doc. Ing. František Drozen, CSc.
- prof. Ing. Václav Kubišta, CSc.
- Ing. Patrik Sieber, PhD.
- prof. RNDr. Hana Skalská, CSc.
- doc. Ing. Jiří Žváček, CSc.

Visegrád, 08/ 06/2011

Joint statement of the V6 Group

on the Communication from the European Commission to the European Parliament and the Council "Towards robust quality management for European Statistics"

V6 is the association of the following six statistical societies:

Austrian Statistical Society
Czech Statistical Society
Hungarian Statistical Association
Romanian Statistical Society
Slovak Statistical and Demographical Society
Statistical Society of Slovenia

The objective of V6 is to facilitate, among others, scientific progress, promote the application of professional ethics and the fundamental principles of official statistics, and best practices in statistics in their respective countries. The members of the V6 group welcome the Communication from the European Commission to the European Parliament and the Council "Towards robust quality management for European Statistics". At the meeting of June, 8th 2011 in Visegrád, Hungary, the paper was discussed and highly appreciated.

The V6 members agree that the principle of professional independence is of core importance for the reliability and credibility of official statistics. To maintain professional independence beyond any reasonable doubt is primarily the task of the respective political systems of the member states and the EU.

The V6 calls for measures to secure this by an explicit commitment.

Austrian Statistical Society

Czech Statistical Society

Hungarian Statistical Association

Romanian Statistical Society

Slovak Statistical and Demographical Society

Statistical Society of Slovenia

Contents / Obsah

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

Časopis je zařazen do seznamu Rady pro výzkum, vývoj a inovace, více viz server `http://www.vyzkum.cz/`

The Bulletin of the Czech Statistical Society is published quarterly. Most of the contributions are published in Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)

Toto číslo bylo vytištěno s laskavou podporou Českého statistického úřadu.