INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 26, číslo 3, září 2015

NELINEÁRNÍ REGRESE V PŘÍKLADECH
NONLINEAR REGRESSION IN EXAMPLES

Karel Zvára

Adresa: ÚAMVT PřF UK v Praze, Albertov 6, 128 43, Praha 2,
KPMS MFF UK, Sokolovská 83, 186 75, Praha 8 – Karlín

E-mail: k@zvara.cz

Abstrakt: Na dvou úlohách článek ukazuje, jak lze pomocí R řešit základní problémy nelineární regrese: odhad parametrů, intervaly spolehlivosti, reparametrizace a míry křivosti.

Klíčová slova: nelineární regrese, interval spolehlivosti, reparametrizace, míry křivosti.

Abstract: On two tasks article demonstrates how to use R to solve basic nonlinear regression problems: estimation of parameters, confidence intervals, reparametrization and curvature measures.

Keywords: nonlinear regression, confidence interval, reparametrization, curvature measures.

1. Úvodem

Nelineární regrese mne doprovází drahná desetiletí. Kdysi jsem dělal pro docenta Janků z Farmakologického ústavu spoustu kompartmentových modelů, také jsem počítal něco pro svoji známou z Ústavu hematologie a krevní transfuse. Konečně, ještě předloni jsem o regresi, i nelineární, přednášel studentům. Tak bych rád doplnil článek brněnských kolegů [2] o další pohled. Vede mne k tomu i zkušenost s balíkem R, bez kterého si už nedovedu svoji práci ani představit. Dovolte mi tedy, abych jako další cvičení udělal tomuto balíku nějakou další reklamu.

2. Bodový a intervalový odhad

V prvním příkladu článku se jedná o závislost počtu bakterií na čase, předpokládá se závislost tvaru $y = \exp(\beta_0 + \beta_1 x)$. Autoři hledají odhad parametrů pomocí postupně zdokonalovaných lineárních aproximací nelineární regresní funkce, přičemž aproximace zlepšují pomocí vah. K odhadu parametrů použijeme tentokrát přímo proceduru `nls()` ze standardní R knihovny `stats`. Musíme zvolit výchozí aproximace odhadů, ale vzhledem k malé nelinearitě úlohy (vrátíme se k ní později) na této volbě příliš nezáleží:

```
> library(stats)
> x <- c(2.5, 2.8, 5.4, 6.5, 9.2, 9.5, 11, 13.3, 14.6, 16.4)
> y <- c(3.03, 6.213, 13.91, 19.305, 27.037, 27.381, 49.845,
+        55.069, 55.453, 75.943)
> a11 <- nls(y~exp(b0+b1*x),start=c(b0=0.5,b1=1))
> summary(a11)

Formula: y ~ exp(b0 + b1 * x)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0  2.01985    0.23152   8.724 2.33e-05 ***
b1  0.14265    0.01644   8.679 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.472 on 8 degrees of freedom

Number of iterations to convergence: 18
Achieved convergence tolerance: 2.522e-06
> deviance(a11)
[1] 335.0784
```

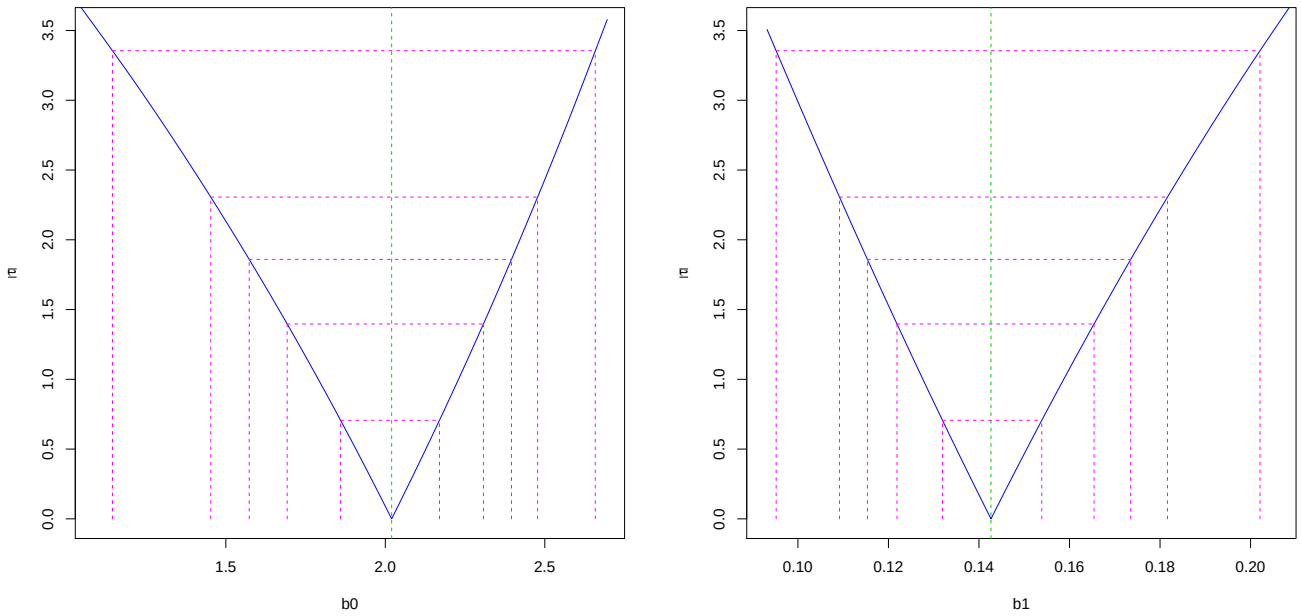
(Připomeňme, že funkce `deviance()` poskytuje v případě lineární či nelineární regrese reziduální součet čtverců.) Skutečnost, že jsme potřebovali 18 iterací není podstatná. V porovnání s článkem [2] jsme dostali výsledné odhady až po 18 iteracích, ale máme o těchto odhadech více informací. Odhady jsou velice podobné, reziduální součet čtverců je v zobrazené přesnosti stejný. Snadno však dostaneme také intervaly spolehlivosti pro oba odhady:

```
> confint(a11)

        2.5%     97.5%
b0 1.4527257 2.4774009
b1 0.1091964 0.1816406
```

Jsou prakticky stejné, k jakým došli autoři článku [2], nejsou však závislé na simulacích. Nejde o primitivní intervalové odhady založené na středních chybách odhadů parametrů, ty bychom dostali například pomocí

```
> SE = summary(a11)$coefficients[,2]
> tKrit = qt(0.975,summary(a11)$df[2])
> CI = cbind(coef(a11),SE)%*%matrix(c(1,-tKrit,1,tKrit),2,2)
> colnames(CI) = c("2.5%","97.5%")
> CI
```



Obrázek 1: Profilové diagramy v původní parametrizaci.

```
        2.5%     97.5%
b0 1.4859564 2.5537454
b1 0.1047473 0.1805511
```

Funkce `confint()` pracuje s výsledkem procedury `nls()` jemněji, používá profilovou funkci [4, str. 195], kterou si pro jednotlivé parametry regresní funkce můžeme nechat nakreslit pomocí dvou příkazů `par(mfrow=c(1,2))` a `plot(profile(a12))`.

Na vodorovné ose jsou hodnoty příslušného parametru, svislá osa se vztahuje k absolutní hodnotě profilové funkce. Čárkovaně lze nalézt přibližné 90% a 95% intervaly spolehlivosti. Zvláště u odhadu parametru β_0 je vidět mírná nelinearita, která vede také k určité nesymetrii intervalů spolehlivosti vzhledem k bodovému odhadu.

2.1. Odhady odvozených parametrů

Pokud nás zajímají spíše odvozené parametry, totiž počet bakterií v čase 0 ($\nu_0 = \exp(\beta_0)$, odhad označíme N_0) a čas, kdy dojde ke zdvojnásobení počtu bakterií (ξ_2, odhad x_2), máme před sebou dvě možnosti. První znamená spočítat odhady a intervaly spolehlivosti na základě aproximací transformací parametrů pomocí Taylorova rozvoje:

```
> SE1 = c(exp(coef(a11)[1])*SE[1],log(2)/coef(a11)[2]^2*SE[2])
> odhad = c(exp(coef(a11)[1]),log(2)/coef(a11)[2])
```

```
> tab = cbind(odhad,odhad-SE1*tKrit,odhad+SE1*tKrit)
> rownames(tab) = c("N0","x2")
> colnames(tab) = c("odhad","2,5 %","95,5 %")
> tab
      odhad    2,5 %    95,5 %
N0 7.537201 3.513131 11.561272
x2 4.859104 3.568040  6.150168
```

2.2. Reparametrizace

Druhou možností je vyjádřit regresní funkci pomocí nových parametrů. Zaveďme parametry

$$\nu_0 = \exp(\beta_0), \quad \xi_2 = \ln(2)/\beta_1.$$

Regresní funkce má pak tvar $y = \nu \exp(\ln(2) \cdot x/\xi)$. Podobně jako při původní parametrizaci dojdeme k odhadům a intervalům spolehlivosti:

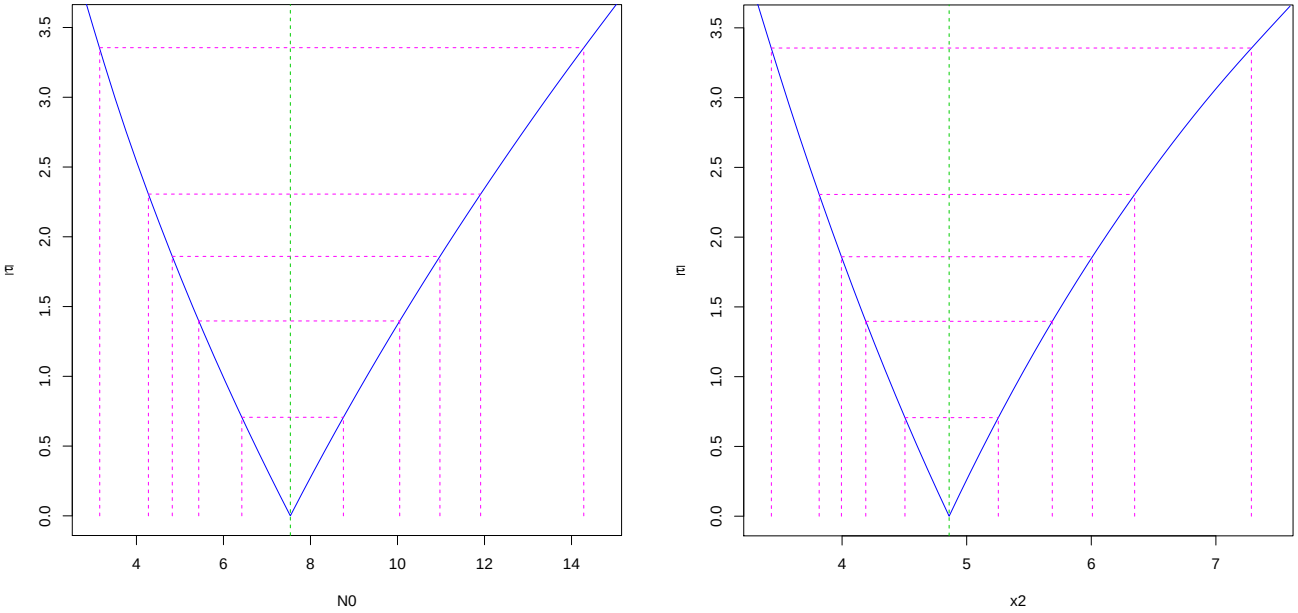
```
> a12 <- nls(y~N0*exp(x*log(2)/x2),start=c(N0=0.5,x2=1))
> summary(a12)

Formula: y ~ N0 * exp(x * log(2)/x2)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
N0   7.5372     1.7450   4.319  0.00255 **
x2   4.8591     0.5599   8.679 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.472 on 8 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 8.214e-06
> deviance(a12)
[1] 335.0784
> confint(a12)
       2.5%     97.5%
N0 4.275679 11.910946
x2 3.816504  6.348343
```

Bodové odhady jsou pochopitelně stejné, jako vyšly přímým výpočtem, stejný je i reziduální součet čtverců. Rozdíl je pouze v intervalech spolehlivosti, které už nejsou symetrické kolem bodových odhadů.



Obrázek 2: Profilové diagramy v upravené parametrizaci.

2.3. Měření nelinearity

Nelinearitu je možné měřit. K odhadu nelinearity potřebujeme pracovat s prvními a druhými parciálními derivacemi regresní funkce podle jejích parametrů. Naštěstí nemusíme derivace ručně odvozovat, zpravidla to za nás R dokáže udělat samo:

```
> fce13 <- deriv3(~exp(b0+b1*x), namevec=c("b0","b1"),
+                 function.arg=function(x,b0,b1){})
> a13 = nls(y~fce13(x,b0,b1),start=c(b0=0,b1=1))
> summary(a13)

Formula: y ~ fce13(x, b0, b1)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0  2.01985    0.23152   8.724 2.33e-05 ***
b1  0.14265    0.01644   8.679 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.472 on 8 degrees of freedom

Number of iterations to convergence: 18
Achieved convergence tolerance: 1.922e-06
```

```
> deviance(a13)
[1] 335.0784
> library(MASS)
> rms.curv(a13)
Parameter effects: c^theta x sqrt(F) = 0.2272
        Intrinsic: c^iota  x sqrt(F) = 0.1218
> confint(a13)
        2.5%     97.5%
b0 1.4527257 2.4774009
b1 0.1091964 0.1816406
```

Procedura `deriv3()` připravila výpočet nejen funkční hodnoty naší regresní funkce, ale také zmíněných prvních a druhých parciálních derivací. Procedura `rms.curv()` počítá průměrnou parametrickou a průměrnou vnitřní křivost [4, str. 211].

R tiskne tyto hodnoty vynásobené výrazem $\sqrt{F_{k,n-2}(0{,}95)}$, kde n je počet pozorování, k počet odhadovaných regresních koeficientů a $F_{k,n-2}(0{,}95)$ 95% kvantil F-rozdělení. To umožňuje porovnávat křivost v různých úlohách s různými hodnotami k, n. Velikost obou křivostí je v naší úloze tolerovatelná. Provedeme-li analogický výpočet pro druhou parametrizaci uvažované závislosti, dostaneme

```
> fce14 <- deriv3(~N0*exp(x*log(2)/x2), namevec=c("N0","x2"),
+                 function.arg=function(x,N0,x2){})
> a14 = nls(y~fce14(x,N0,x2),start=c(N0=0.5,x2=1))
> summary(a14)

Formula: y ~ fce14(x, N0, x2)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
N0   7.5372     1.7450   4.319  0.00255 **
x2   4.8591     0.5599   8.679 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.472 on 8 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 8.218e-06
> deviance(a14)
```

```
[1] 335.0784
> rms.curv(a14)
Parameter effects: c^theta x sqrt(F) = 0.584
        Intrinsic: c^iota  x sqrt(F) = 0.1218
> confint(a14)
       2.5%     97.5%
N0 4.275679 11.910946
x2 3.816504  6.348343
```

Vnitřní křivost závisí na tvaru regresní funkce, nezmění se, když stejnou závislost y na x vyjádříme pomocí jiné sady parametrů. Na parametrickém vyjádření však závisí parametrická křivost, která je tentokrát větší. To se projeví i na aproximaci vychýlení bodových odhadů. (Připomeňme, že v nelineární regresi odhady parametrů už nemusí být nestranné, jak je tomu v regresi lineární.) Knihovní funkce `rms.curv` však informaci o odhadu vychýlení neposkytuje, takže použijeme modifikaci uvedenou v [4, odst. A.4.7].

```
> source("MODIFrms.R")
> Mrms.curv(a13)

Parameter effects:                0.1076  (max 0.1558 )
        Intrinsic:                0.0577  (max 0.0942 )
Parameter effects (x sqrt(F)):    0.2272  (max 0.3289 )
        Intrinsic (x sqrt(F)):    0.1218  (max 0.1989 )
         Estimate      Bias   Rel. Bias
b0           2.02 -0.010619    -0.526 %
b1          0.143  0.000584     0.409 %
> Mrms.curv(a14)
Parameter effects:                0.2766  (max 0.2729 )
        Intrinsic:                0.0577  (max 0.0942 )
Parameter effects (x sqrt(F)):     0.584  (max 0.5763 )
        Intrinsic (x sqrt(F)):    0.1218  (max 0.1989 )
         Estimate      Bias   Rel. Bias
N0          7.537  0.121975     1.618 %
x2          4.859  0.044627     0.918 %
```

Při různé parametrizaci modelu lze porovnávat jen relativní vychýlení (vychýlení vztažené k bodovému odhadu), které v našem případě při nové parametrizaci závislosti poněkud vzrostlo. Čtenář jistě zaznamenal, že kromě průměrných křivostí počítá modifikovaná procedura také maximální křivosti, které podle některých autorů dávají lepší představu o skutečnosti.

2.4. R najde výchozí aproximaci

Až dosud jsme při výpočtu odhadů museli volit výchozí aproximaci. Pro řadu užívaných regresních závislostí má R připraveny funkce, které tuto výchozí aproximaci připraví samy. Jejich názvy začínají písmeny SS, například `SSlogist, SSmicmen, SSweibull`. Je tu také připraven postup, jak si sestavit vlastní funkci se „samostartem“. Podstatná je část `initial`, kde je sestrojena funkce, která dokáže z volání funkce `SSinv()` najít hodnoty nezávisle i závisle proměnné a spočítá výchozí aproximaci odhadu parametrů stejně, jak to učinili autoři článku [2] v prvním kroku, bez použití váhové funkce. Netvrdím, že jsem tuto funkci sestavil ideálně, bylo by například vhodné ohlídat, zda jsou hodnoty vysvětlované proměnné „dostatečně kladné“.

```
> # příprava samostartující funkce
> SSinv = selfStart(~exp(b0+b1*x),
+   initial = function(mCall, data, LHS){
+     y = eval(LHS,data)
+     z = log(y)
+     x = eval(mCall[["x"]],data)
+     aa = lm(z~x,data=list(x=x,z=z))
+     value = c(coef(aa)[1],coef(aa)[2])
+     names(value) = mCall[c("b0","b1")]
+     return(value)
+   },
+   parameters = c("b0","b1"),
+   template = function(x,b0,b1){}
+ )
```

Vlastní odhad je velice jednoduchý. Abychom viděli jednotlivé aproximace odhadu, zejména aproximaci výchozí, volbou `trace=TRUE` požádáme o výpis jednotlivých iterací:

```
> summary(nls(y~SSinv(x,b0,b1),trace=TRUE))

1271.514 :  1.272561 0.204744
356.8708 :  2.142265 0.135745
335.1314 :  2.0134945 0.1431612
335.0791 :  2.0207186 0.1425862
335.0784 :  2.0197468 0.1426568
335.0784 :  2.0198653 0.1426481
335.0784 :  2.0198508 0.1426492
Formula: y ~ SSinv(x, b0, b1)
```

```
Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0  2.01985    0.23152   8.724 2.33e-05 ***
b1  0.14265    0.01644   8.679 2.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.472 on 8 degrees of freedom

Number of iterations to convergence: 6
Achieved convergence tolerance: 2.774e-06
```

2.5. Druhý příklad

Na rozdíl od prvního příkladu z článku [2], kdy se jednalo o proces množení, tentokrát jde o proces hynutí. Pokles počtu bakterií po aplikaci antibiotika je dán vztahem $y = 1/(\beta_0 + \beta_1 x)$, kde y je počet bakterií v jednotce objemu a x je čas. Cílem má být odhadnout čas, kdy dojde k poklesu počtu bakterií na polovinu. Odhad parametrů pomocí standardní funkce `nls()` vyjde samozřejmě stejně, jako v článku [2]:

```
> x <- c(0,0.5,0.7,1,1.3,1.5,2,2.5,3,3.5,4,4.5,5,5.5,6)
> y <- c(51.3, 31.74, 23, 15.99, 8, 4.81, 4.25, 2.19, 0.25,
+        2.23, 0.2, 1.19, 0.18, 0.31, 0.25)
> a2 = nls(y~1/(b0+b1*x^2),start=c(b0=1,b1=1))
> summary(a2)

Formula: y ~ 1/(b0 + b1 * x^2)

Parameters:
    Estimate Std. Error t value Pr(>|t|)
b0 0.0193488  0.0004413   43.84 1.64e-15 ***
b1 0.0514099  0.0027898   18.43 1.06e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 13 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 8.102e-07
```

```
> deviance(a2)
[1] 19.07035
```

Nás ale zajímají spíše odvozené parametry, totiž počet bakterií v čase 0 (parametr označíme ν_0, odhad pak N_0) a čas, kdy dojde k polovičnímu počtu bakterií (parametr ξ_2, odhad x_2).

Abychom mohli určit také míru nelinearity úlohy, připravíme si opět výpočet regresní funkce včetně jejích prvních a druhých parciálních derivací:

```
> fce21 = deriv3(~1/(b0+b1*x*x),namevec=c("b0","b1"),
+                function.arg=function(x,b0,b1){})
> a21 = nls(y~fce21(x,b0,b1),start=c(b0=1,b1=1))
> summary(a21)

Formula: y ~ fce21(x, b0, b1)

Parameters:
    Estimate Std. Error t value Pr(>|t|)
b0 0.0193488  0.0004413   43.84 1.64e-15 ***
b1 0.0514099  0.0027898   18.43 1.06e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 13 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 8.099e-07
> deviance(a21)
[1] 19.07035
> Mrms.curv(a21)
Parameter effects:                0.0672  (max 0.091 )
        Intrinsic:                0.0245  (max 0.0401 )
Parameter effects (x sqrt(F)):    0.1312  (max 0.1775 )
        Intrinsic (x sqrt(F)):    0.0479  (max 0.0782 )
         Estimate      Bias   Rel. Bias
b0          0.019     9e-06     0.045 %
b1          0.051   7.1e-05     0.139 %
> confint(a21)
         2.5%      97.5%
b0 0.01844556 0.02033899
b1 0.04585718 0.05769751
```

Provedeme-li analogický výpočet pro druhou parametrizaci uvažované závislosti, dostaneme

```
> fce22 = deriv3(~c1*c2*c2/(c2*c2+x*x),namevec=c("c1","c2"),
+                function.arg=function(x,c1,c2){})
> a22 = nls(y2~fce22(x2,c1,c2),start=c(c1=1,c2=5))
> summary(a22)

Formula: y ~ fce22(x, c1, c2)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
c1 51.68273    1.17889   43.84 1.64e-15 ***
c2  0.61349    0.01999   30.68 1.63e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.211 on 13 degrees of freedom

Number of iterations to convergence: 14
Achieved convergence tolerance: 7.535e-06
> deviance(a22)
[1] 19.07035
> Mrms.curv(a22)
Parameter effects:               0.0545  (max 0.092 )
        Intrinsic:               0.0245  (max 0.0401 )
Parameter effects (x sqrt(F)):   0.1063  (max 0.1795 )
        Intrinsic (x sqrt(F)):   0.0479  (max 0.0782 )
         Estimate      Bias   Rel. Bias
c1         51.683   0.00382     0.007 %
c2          0.613  0.000408     0.067 %
> confint(a22)
         2.5%      97.5%
c1 49.1668203 54.2136496
c2  0.5730101  0.6575098
```

Z výstupů je patrné, že průměrná parametrická křivost je u druhé parametrizace nepatrně menší než u parametrizace původní. Maximální parametrická křivost je v obou případech prakticky stejná. Rozdíl je patrný u relativního vychýlení, které je u číselně větších odhadů nových regresních koeficientů menší.

2.6. Poznámka

Je otázka, nakolik jsou splněny běžné předpoklady modelu, zejména předpoklad homoskedasticity. U funkce $y = 1/(\beta_0 + \beta_1 x)$ z druhého příkladu bych se obával, že rozptyl bude s rostoucím časem x klesat. Proto jsem zkusil vyšetřit závislost druhých mocnin reziduí na vyrovnaných hodnotách, ale závislost jsem neprokázal:

```
anova(lm(residuals(a22)^2~predict(a22)))
```

```
Analysis of Variance Table
Response: residuals(a22)^2
             Df Sum Sq Mean Sq F value Pr(>F)
predict(a22)  1  0.404  0.4045  0.1057 0.7502
Residuals    13 49.722  3.8247
```

Testová statistika zhruba odpovídá testu homoskedasticity, který navrhli Breusch a Pagan ([1], viz též [4, str. 128]). Podobně dopadlo hodnocení reziduí pomocí Shapirova-Wilkova testu normality:

```
shapiro.test(resid(a22))
```

```
        Shapiro-Wilk normality test
data:  resid(a22)
W = 0.9827, p-value = 0.9846
```

Poděkování

Děkuji dvěma anonymním recenzentům za podnětné připomínky.

Literatura

[1] Breusch T. S., Pagan A. R.: A simple test for heteroscedasticity and random coefficient variation. *Econometrica* **47**, 1287–1294, 1979.

[2] Maroš B., Budíková M.: Význam váhové funkce v linearizovatelných regresních modelech. *Informační bulletin České statistické společnosti* **12**, 1–9, 2012.

[3] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[4] Zvára K.: *Regrese*. MATFYZPRESS, Praha, 2008. Dostupné též na adrese `http://www.mff.cuni.cz/fakulta/mfp/download/books/zvara_-_regrese.pdf`, procedury jsou dostupné z `http://www.karlin.mff.cuni.cz/~zvara/regrese/kniha/kniha08.html`.

LINEÁRNÍ SMÍŠENÉ REGRESNÍ MODELY PŘI SLEDOVÁNÍ OBSAHU TĚŽKÝCH KOVŮ V SEDIMENTECH ŘEKY MORAVY

LINEAR MIXED MODELS IN MONITORING HEAVY METALS CONTENT OF THE MORAVA RIVER SEDIMENTS

Marie Forbelská[1], **Hana Hudcová**[2], **Ilja Bernardová**[2], **Jana Svobodová**[2]

Adresa: [1] Ústav matematiky a statistiky, Přírodovědecká fakulta Masarykovy univerzity, Kotlářská 2, 611 37 Brno,
[2] Výzkumný ústav vodohospodářský T. G. Masaryka, v.v.i., Pobočka Brno, Mojmírovo náměstí 16, 612 00 Brno

E-mail: forbel@math.muni.cz[1], Hana_Hudcova@vuv.cz[2]

Abstrakt: Lineární smíšené modely (LMM modely), které jsou velmi často využívány při analýze skupinově závislých dat, byly použity při sledování dlouhodobého vývoje obsahu prioritních a dalších nebezpečných látek v sedimentech. V sedmi lokalitách, situovaných v podélném profilu řeky Moravy mezi 298. a 93. říčním kilometrem, byly sledovány čtyři ukazatele ze skupiny těžké kovy – kadmium, olovo, rtuť a nikl. Příspěvek popisuje výsledky účelového sledování a navazujícího statistického hodnocení časového vývoje obsahu těžkých kovů v sedimentech řeky Moravy v letech 1997–2010.

Klíčová slova: lineární smíšené modely, LMM, nebezpečné látky v sedimentech, časový vývoj, řeka Morava.

Abstract: Linear mixed models (LMM) are frequently used in the analysis of group-dependent data, and were therefore used to monitor long-term trends of priority and other hazardous substances in sediments. In seven sites, located in the longitudinal section of the Morava River between 298–93 rkm, four indicators were monitored from the group of heavy metals – cadmium, lead, mercury and nickel. This paper presents the first results of special monitoring and subsequent statistical evaluation of long-term trends of selected heavy metals content in the Morava River sediments in 1997–2010.

Keywords: Linear mixed models, LMM, hazardous substances in sediments, long-term trends, Morava River.

1. Úvod

Míra znečištění sedimentů prioritními a dalšími nebezpečnými látkami představuje pro své potenciálně toxické účinky na faunu a flóru dna i vodního sloupce nad sedimentem, včetně schopnosti akumulace v tělech vodních živočichů, jeden z hlavních environmentálních problémů. V rámci národních výzkumných monitorovacích projektů ochrany vod – projektů „Morava I – IV“ (viz [1]–[3], [13]) a projektu „Identifikace antropogenních tlaků na kvalitativní stav vod a vodních ekosystémů v oblasti povodí Moravy“ (viz [5]) v letech 1997 až 2010 probíhalo hodnocení kvalitativního stavu sedimentů nejvíce zatížených úseků řeky Moravy. Právě prověření časového vývoje obsahu nebezpečných látek v sedimentech je jedním z aktuálních požadavků směrnice Evropského parlamentu a Rady 2008/105/ES o normách environmální kvality (viz [10]).

Z hlediska modelování trendů zátěže sedimentů toků pod významnými zdroji znečištění těžkými kovy z let 1997–2010 byly pro sledované lokality na řece Moravě využity stochastické modely, které se dokáží vyrovnat jak s heterogenními rozptyly, tak s korelovanými daty.

2. Matematická formulace problému pomocí lineárních smíšených regresních modelů

Klasické statistické metody se obvykle zajímají buď o nezávislá pozorování nebo o časové řady. V tomto případě však máme k dispozici data, která se týkají několika lokalit, ve kterých jsou opakovaně získávána závislá pozorování sledovaná v čase. Ke správné a efektivní analýze takových dat použijeme lineární smíšené modely, které umožňují zohlednit vliv konkrétních lokalit na svoje opakovaná měření. Vložením individuálních efektů jednotlivých lokalit lze odhadovat nejen celkovou změnu společnou všem, ale také specifické změny každé lokality.

2.1. Značení a tvorba LMM modelu

Při modelování čtrnáctiletého vývoje obsahu nebezpečných látek v sedimentech měřeného v sedmi lokalitách vyjdeme z jednoduchého růstového modelu. Proto označme indexem j $(j = 1, \ldots, N)$ jednotlivé lokality, dvojicemi indexů ji $(i = 1, \ldots, n_j)$ korelovaná pozorování na j-té lokalitě a symbolem $n = \sum_{j=1}^{N} n_j$ celkový počet pozorování a uvažujme *lineární růstový model*

$$Y_{ji} = \beta_0 + \beta_1 t_{ji} + \varepsilon_{ji},$$

kde t_{ji} je čas i-tého pozorování lokality j a ε_{ji} jsou nekorelované náhodné odchylky s nulovou střední hodnotou a konstantním rozptylem. Za těchto předpokladů

$$\mathrm{E}\, Y_{ji} = \beta_0 + \beta_1 t_{ji}.$$

Pokud například odezvu Y_{ji} interpretujeme ve vztahu ke zpracovávaným datům jako logaritmus obsahu olova ve vzorku sedimentu v čase t_{ji}, pak parametr β_0 označuje průměrnou hodnotu logaritmu obsahu olova ve vzorku v čase 0 a β_1 je jeho průměrný přírůstek za jednotku času.

Předpokládejme nyní, že logaritmy obsahu olova ve stejné lokalitě mají tendenci být buď systematicky vyšší nebo systematicky nižší než průměr, a to po celou dobu sledování.

Proto pro j-tou lokalitu zavedeme nepozorovanou normálně rozdělenou náhodnou odchylku od průměru $b_{j0} \sim N(0, \sigma^2_{b_0})$ a model upravíme takto

$$Y_{ji} = \beta_0 + \beta_1 t_{ji} + \underbrace{b_{j0} + \eta_{ji}}_{\varepsilon_{ji}} \tag{1}$$

kde $\eta_{ji} \sim \text{iid}\, N(0, \sigma^2_j)$ a η_{ji} s b_{j0} jsou nezávislé. Pak

$$\begin{aligned}\rho_j = \mathrm{corr}(Y_{ji}, Y_{ji'}) = \mathrm{corr}(\varepsilon_{ji}, \varepsilon_{ji'}) &= \frac{\mathrm{cov}(\varepsilon_{ji}, \varepsilon_{ji'})}{\sqrt{\mathrm{var}(\varepsilon_{ji})}\sqrt{\mathrm{var}(\varepsilon_{ji'})}} \\ &= \frac{\mathrm{var}(b_{j0})}{\mathrm{var}(b_{j0}) + \mathrm{var}(\eta_{ji})} = \frac{\sigma^2_{b_0}}{\sigma^2_{b_0} + \sigma^2_j}.\end{aligned}$$

Navíc pořád bude platit

$$\mathrm{E}\, Y_{ji} = \beta_0 + \beta_1 t_{ji}.$$

Nyní budeme předpokládat, že logaritmy obsahu olova v sedimentech v sledovaných lokalitách se liší nejen polohou, ale také směrnicí růstu. Zavedeme proto další nepozorovanou normálně rozdělenou náhodnou veličinu $b_{j1} \sim N(0, \sigma^2_1)$ a model obdobně upravíme

$$Y_{ji} = \beta_0 + \beta_1 t_{ji} + \underbrace{b_{j0} + b_{j1} t_{ji} + \eta_{ji}}_{\varepsilon_{ji}}, \tag{2}$$

kde η_{ji} a $(b_{j0}, b_{j1})'$ jsou nezávislé a $\mathrm{cov}(b_{j0}, b_{j1}) = \sigma_{b_0 b_1}$. Pak

$$\begin{aligned}\rho_j(t_{ji}, t_{ji'}) = \mathrm{corr}(Y_{ji}, Y_{ji'}) = \mathrm{corr}(\varepsilon_{ji}, \varepsilon_{ji'}) &= \frac{\mathrm{cov}(\varepsilon_{ji}, \varepsilon_{ji'})}{\sqrt{\mathrm{var}(\varepsilon_{ji})}\sqrt{\mathrm{var}(\varepsilon_{ji'})}} \\ &= \frac{\sigma^2_{b_0} + \sigma_{b_0 b_1}(t_{ji} + t_{ji'}) + \sigma^2_{b_1} t_{ji} t_{ji'}}{\sqrt{\sigma^2_{b_0} + 2\sigma_{b_0 b_1} t_{ji} + \sigma^2_{b_1} t^2_{ji} + \sigma^2_j}\sqrt{\sigma^2_{b_0} + 2\sigma_{b_0 b_1} t_{ji'} + \sigma^2_{b_1} t^2_{ji'} + \sigma^2_j}}\end{aligned}$$

a vidíme, že v modelu s náhodným posunutím i náhodnou směrnicí jak rozptyl Y_{ji}, tak kovariance mezi Y_{ji} a $Y_{ji'}$ se mění s časem.

Obecný model s náhodnými i nenáhodnými (pevnými) efekty (proto název lineární smíšený model) zobecňuje princip, který jsme ilustrovali zavedením náhodného absolutního členu a náhodné směrnice.

Pokud položíme $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jn_j})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $\mathbf{b}_j = (b_{j0}, b_{j1})'$ a vytvoříme matice plánu $\mathbf{X}_j$ a $\mathbf{Z}_j$ s n_j řádky ve tvaru

$$\mathbf{x}_{ji} = \mathbf{z}_{ji} = (1, t_{ji})' \quad (i = 1, \ldots, n_j),$$

můžeme pro každou lokalitu j vztahy (2) vyjádřit pomocí maticového zápisu

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \underbrace{\mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\eta}_j}_{\boldsymbol{\varepsilon}_j}, \tag{3}$$

kde matice plánu $\mathbf{X}_j$ je typu $n_j \times k$ ($k = 2$) a $\mathbf{Z}_j$ je typu $n_j \times q$ ($q = 2$), vektor pevných efektů $\boldsymbol{\beta}$ je typu $k \times 1$, vektor náhodných efektů $\mathbf{b}_j \sim N_q(\mathbf{0}, \mathbf{D})$ je typu $q \times 1$, přitom platí $\boldsymbol{\eta}_j \sim N_{n_j}(\mathbf{0}, \boldsymbol{\Sigma}_j)$ a oba normální vektory $\mathbf{b}_j$ a $\boldsymbol{\eta}_j$ jsou nezávislé.

Zavedením náhodných efektů modelujeme varianční strukturu vektoru $\mathbf{Y}_j$

$$\mathbf{V}_j = \mathbf{V}(\boldsymbol{\psi}_j) = \mathrm{var}(\mathbf{Y}_j) = \mathrm{var}(\boldsymbol{\varepsilon}_j) = \mathrm{var}(\mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\eta}_j) = \mathbf{Z}_j\mathbf{D}\mathbf{Z}_j' + \boldsymbol{\Sigma}_j$$

a tím i korelace mezi Y_{ji} a $Y_{ji'}$. Matice $\mathbf{V}_j$ obsahuje neznámé varianční komponenty $\boldsymbol{\psi}_j$, v našem případě σ_j^2, $\sigma_{b_0}^2$, $\sigma_{b_1}^2$ a $\sigma_{b_0 b_1}$.

I když nás především zajímají odhady neznámých pevných efektů $\boldsymbol{\beta}$ a predikce náhodných efektů $\mathbf{b}_j$, je třeba také provést odhad neznámých variačních komponent $\boldsymbol{\psi} = (\boldsymbol{\psi}_1', \ldots, \boldsymbol{\psi}_N')'$, což se obvykle provádí pomocí metody maximální věrohodnosti (ML-odhad) nebo pomocí její modifikované verze REML (Restricted Maximum Likelihood).

K získání odhadu neznámého vektoru pevných efektů $\boldsymbol{\beta}$ se používají tzv. *empirické nejlepší nestranné odhady*

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{N} \mathbf{X}_j' \mathbf{V}(\widehat{\boldsymbol{\psi}}_j)^{-1} \mathbf{X}_j \right)^{-1} \sum_{j=1}^{N} \mathbf{X}_j' \mathbf{V}(\widehat{\boldsymbol{\psi}}_j)^{-1} \mathbf{Y}_j \tag{4}$$

a pro predikci náhodných efektů $\mathbf{b}_j$ tzv. *empirické nejlepší nestranné lineární prediktory.*

$$\widehat{\mathbf{b}}_j = \mathbf{D}(\widehat{\boldsymbol{\psi}}_j)\mathbf{Z}_j' \mathbf{V}(\widehat{\boldsymbol{\psi}}_j)^{-1} (\mathbf{Y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}). \tag{5}$$

V obou případech $\widehat{\boldsymbol{\psi}}_j$ jsou $\widehat{\boldsymbol{\psi}}_j^{\mathrm{ML}}$ nebo $\widehat{\boldsymbol{\psi}}_j^{\mathrm{REML}}$ odhady (podrobnosti lze najít například v [6], kap. 6 a 9).

Pro testování pevných efektů, náhodných efektů a variančních komponent se využívají testy poměrem věrohodností nebo Waldovy testy (více [12], [4]).

3. Pilotní statistické hodnocení dlouhodobého sledování obsahu vybraných těžkých kovů v sedimentech řeky Moravy pomocí LMM modelů

Výsledky dlouhodobého sledování kvalitativního stavu sedimentů jsou podkladem pro zhodnocení vývoje stavu sledovaných lokalit v podélném profilu řeky Moravy včetně vymezení problematických ukazatelů signalizujících vzestupný trend koncentračních hodnot prioritních a dalších nebezpečných látek.

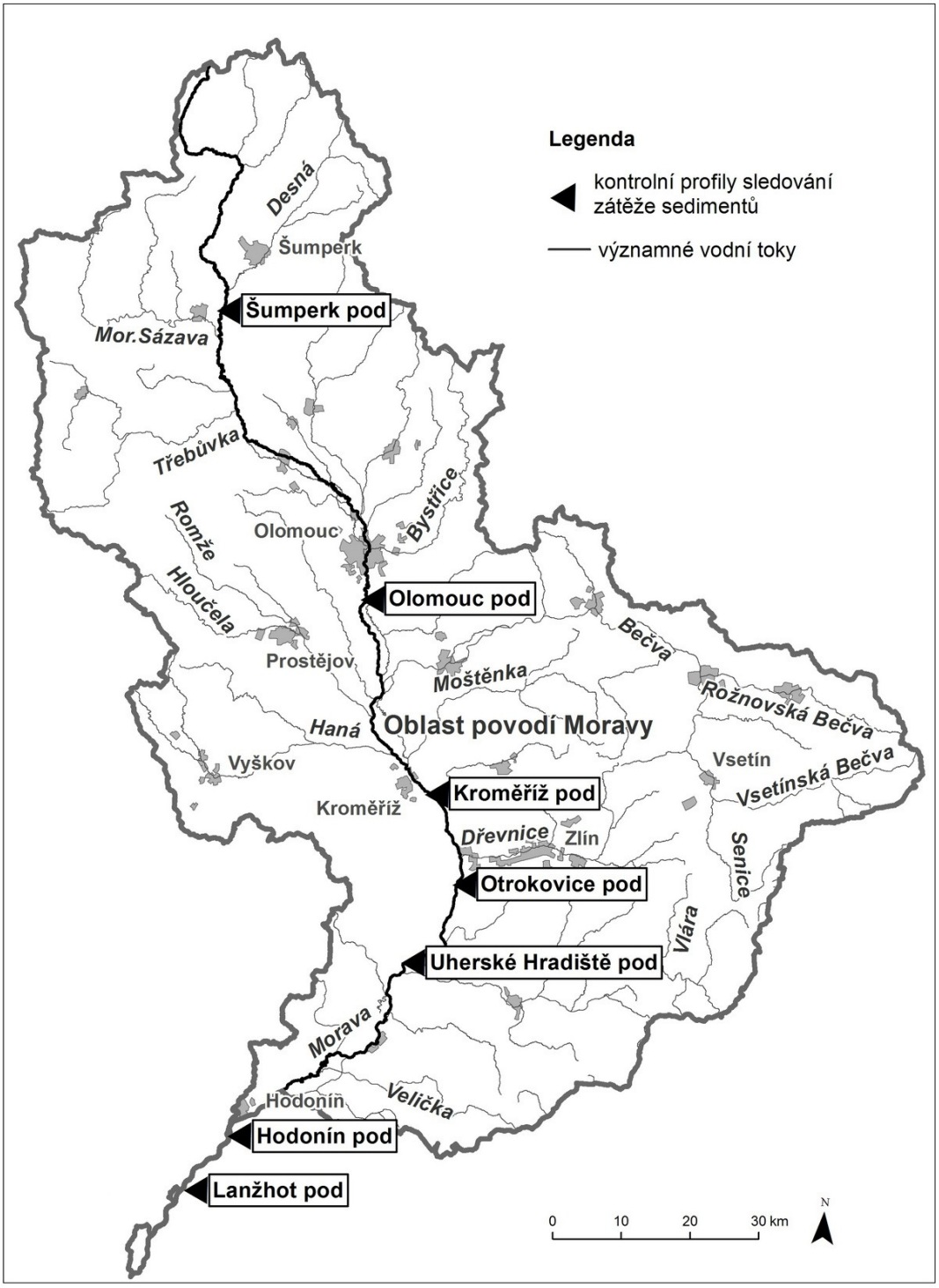
V sedmi lokalitách, situovaných v podélném profilu řeky Moravy mezi 298. a 83. říčním kilometrem, byly sledovány čtyři ukazatele ze skupiny těžké kovy – kadmium, olovo, rtuť a nikl. Jednotlivé profily dlouhodobého sledování zátěže sedimentů jsou znázorněny na obrázku 1 a počty odběrů jsou shrnuty v tabulce 1.

Tabulka 1: Počty odběrů vzorků sedimentů.

Lokalita	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2008	2009	2010	n_j
Šumperk pod	2	2	2	2	2	2	2	2	2	1	1	2	1	23
Olomouc pod	2	2	2	2	2	2	1	2	2	1	2	2	1	23
Kroměříž pod	0	2	2	2	2	2	2	2	2	1	2	2	1	22
Otrokovice pod	2	2	2	2	2	2	1	3	2	1	2	2	1	24
Uherské Hradiště pod	2	2	2	2	2	2	2	2	2	1	2	2	1	24
Hodonín pod	2	2	2	2	2	2	2	2	3	1	0	0	0	20
Lanžhot pod	2	2	0	0	2	2	0	2	2	2	2	2	1	19
Celkem	12	14	12	12	14	14	10	15	15	8	11	12	6	155

Četnost odběru vzorků sedimentů byla dva vzorky ročně s výjimkou roku 2006, kdy byl sediment odebrán pouze jednou. Podobná situace se vzhledem k nepříznivým hydrologickým podmínkám opakovala v roce 2010, kdy byly vzorky sedimentů na všech sledovaných profilech, mimo profil „Otrokovice pod“, odebrány také pouze jednou. V roce 2007 bylo sledování zatížení sedimentů z finančních důvodů na jeden rok přerušeno. Sledování profilu „Hodonín pod“, které probíhalo v letech 1997–2006, muselo být v roce 2008 z organizačních důvodů přesunuto cca 15 km níže po toku pod obec Lanžhot (profil „Lanžhot pod“). Pro komplexní hodnocení zatížení sedimentů řeky Moravy včetně dolního úseku byly do statistického hodnocení zahrnuty oba tyto profily.

Při hodnocení vývoje kvalitativního stavu sedimentů nejvíce zatížených úseků řeky Moravy v letech 1997–2010 jsme nejprve provedli odhad jak lineárního růstového modelu (2) s náhodným posunutím i náhodnou směrnicí, tak lineárního modelu (1) pouze s náhodným posunutím. Ke zjištění, zda



Obrázek 1: Kontrolní profily dlouhodobého sledování zátěže sedimentů.

vystačíme s jednodušším modelem, což je ekvivalentní s testem hypotézy

$$H_0 : \mathbf{D} = \begin{pmatrix} \sigma_{b_0}^2 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{vůči alternativě} \quad H_A : \mathbf{D} = \begin{pmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ \sigma_{b_0 b_1} & \sigma_{b_1}^2 \end{pmatrix},$$

jsme použili test poměrem věrohodností $\mathrm{LR} = -2(l^{\mathrm{REML}}(\widehat{\boldsymbol{\psi}}_0) - l^{\mathrm{REML}}(\widehat{\boldsymbol{\psi}}_A))$, kdy rozdělení LR statistiky bylo aproximováno pomocí směsí dvou χ^2 rozdělení (více viz [11], [7]) a p-hodnota testu byla vypočtena podle vzorce

$$p = 0.5P(\chi_1^2 > \mathrm{LR}_{\mathrm{obs}}) + 0.5P(\chi_2^2 > \mathrm{LR}_{\mathrm{obs}}),$$

kde l^{REML} je logaritmus věrohodnostní funkce v příslušném REML odhadu, $\mathrm{LR}_{\mathrm{obs}}$ je pozorovaná hodnota LR statistiky a χ_k^2 je náhodná veličina s χ^2 rozdělením s k stupni volnosti. Pro všechny čtyři vybrané kovy se ukázalo (viz tabulka 2), že ani v jednom případě nedošlo k významnému zhoršení modelu, když se náhodná směrnice neuvažovala.

Tabulka 2: Výsledné p-hodnoty LR testu.

Pb	Cd	Hg	Ni
0.621	0.989	0.297	0.984

Pro jednotlivé kovy byl pro celou řeku Moravu vytvořen jediný stochastický model, který v sobě zahrnuje regresní přímky jednotlivých lokalit ve tvaru $(\beta_0 + b_{j0}) + \beta_1 t$ (s náhodným posunutím) a také regresní přímku $\beta_0 + \beta_1 t$ pouze s nenáhodnými (pevnými) koeficienty, která popisuje společný vývoj obsahu rizikových látek v sedimentech pro celou řeku Moravu.

Odhady pevných parametrů, které jsou uvedeny v následující tabulce, poslouží k analýze celkového trendu vývoje uvažovaných rizikových látek pro řeku Moravu.

Tabulka 3: Odhady pevných efektů.

Pevné efekty		Odhad	Stand. chyba	St. volnosti	t-stat.	p-hodnota
Pb	β_0	3.5827	0.0572	141	62.6308	<0.0001
	β_1	0.0141	0.0070	141	4.6683	0.0467
Cd	β_0	−0.3182	0.0620	141	−5.1346	<0.0001
	β_1	−0.0246	0.0070	141	−3.5248	0.0006
Hg	β_0	−1.5327	0.0848	137	−18.0771	<0.0001
	β_1	−0.0009	0.0069	137	−0.1281	0.8983
Ni	β_0	3.6599	0.0340	97	107.8190	<0.0001
	β_1	0.0395	0.0076	97	5.1953	<0.0001

Při analýze celkového trendu $\beta_0 + \beta_1 t$ logaritmu hodnot obsahu těžkých kovů v sedimentech řeky Moravy z let 1997 až 2010 byla testována na základě Waldova testu významnost pevného koeficientu β_1. Znaménko parametru β_1 určuje rostoucí ($\beta_1 > 0$), popř. klesající trend ($\beta_1 < 0$). Jeho statistickou významnost udává p-hodnota uvedená v posledním sloupci tabulky 3 a vztahuje se k testování hypotézy $\beta_1 = 0$.

V daném případě byla zamítnuta hypotéza $\beta_1 = 0$ u kadmia (přímka klesá), niklu (přímka mírně stoupá) a olova (těsné zamítnutí, přímka velmi mírně stoupá). Pouze u rtuti se nezamítla hypotéza, že $\beta_1 = 0$.

Statistické hodnocení dat z let 1997–2010 tedy prokázalo klesající trend kadmia, mírně rostoucí trend niklu a minimálně rostoucí trend olova v sedimentech řeky Moravy. Množství rtuti v sedimentech řeky Moravy se ve sledovaném období řádově nezměnilo.

Protože se variabilita kolem trendu v jednotlivých lokalitách výrazně lišila, byl uvažován model, ve kterém varianční matice $\mathbf{\Sigma}_j$ je ve tvaru

$$\mathbf{\Sigma}_j = \sigma_j^2 \mathbf{I}_{n_j} \quad (j = 1, \ldots, 7).$$

Výsledné REML odhady směrodatných odchylek $\sigma_1, \ldots, \sigma_7$ a σ_{b_0} jsou uvedeny v následující tabulce.

Tabulka 4: Varianční komponenty modelu.

Lokalita	σ_j pro Pb	σ_j pro Cd	σ_j pro Hg	σ_j pro Ni
Šumperk pod	0.7153	0.5715	0.9465	0.4539
Olomouc pod	0.2973	0.3646	0.5357	0.3340
Kroměříž pod	0.6157	0.3162	0.3944	0.3243
Otrokovice pod	0.2832	0.3161	0.2787	0.3409
Uherské Hradiště pod	0.4243	0.6844	0.4657	0.3632
Hodonín pod	0.4789	0.6307	0.4744	0.3114
Lanžhot pod	0.2074	0.1869	0.1678	0.3292
Směrodatná odchylka posunutí σ_{b_0}	0.1215	0.1354	0.1997	0.0039

Na základě odhadů variančních komponent lze říci, že oproti ostatním kovům mají hodnoty logaritmu obsahu niklu v sedimentech méně rozdílnou variabilitu kolem trendu a navíc variabilita náhodného posunutí u niklu je zhruba stokrát nižší než u ostatních těžkých kovů.

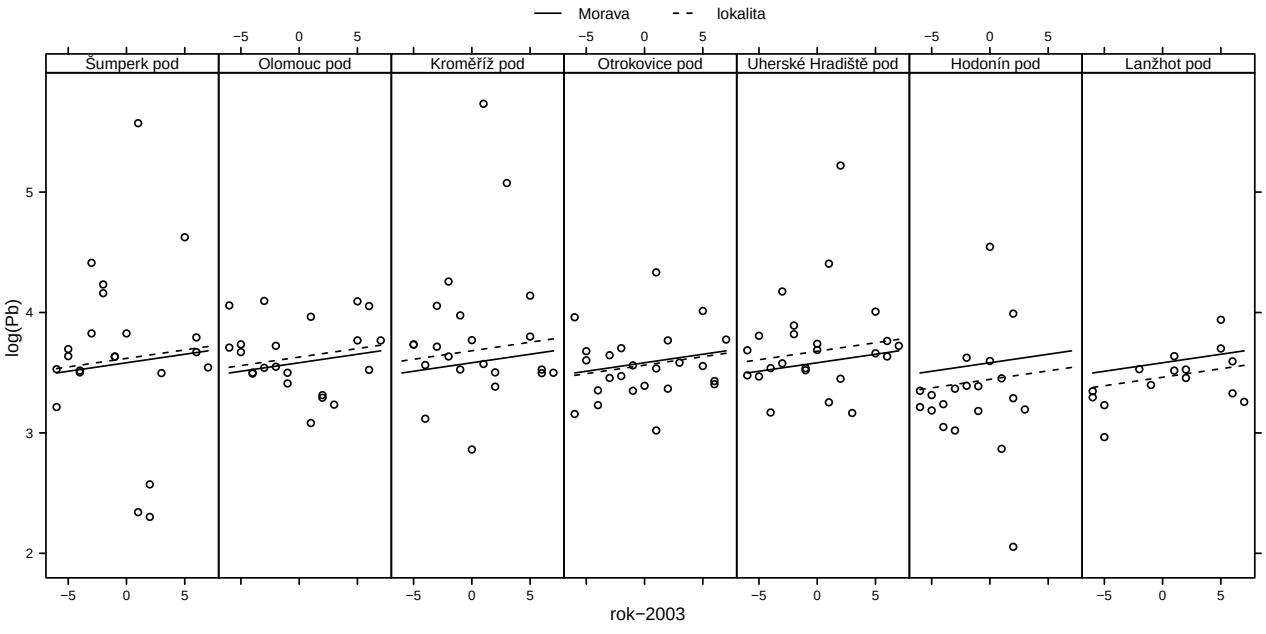
V poslední tabulce jsou uvedeny predikce náhodných posunutí.

Tabulka 5: Predikce náhodných posunutí.

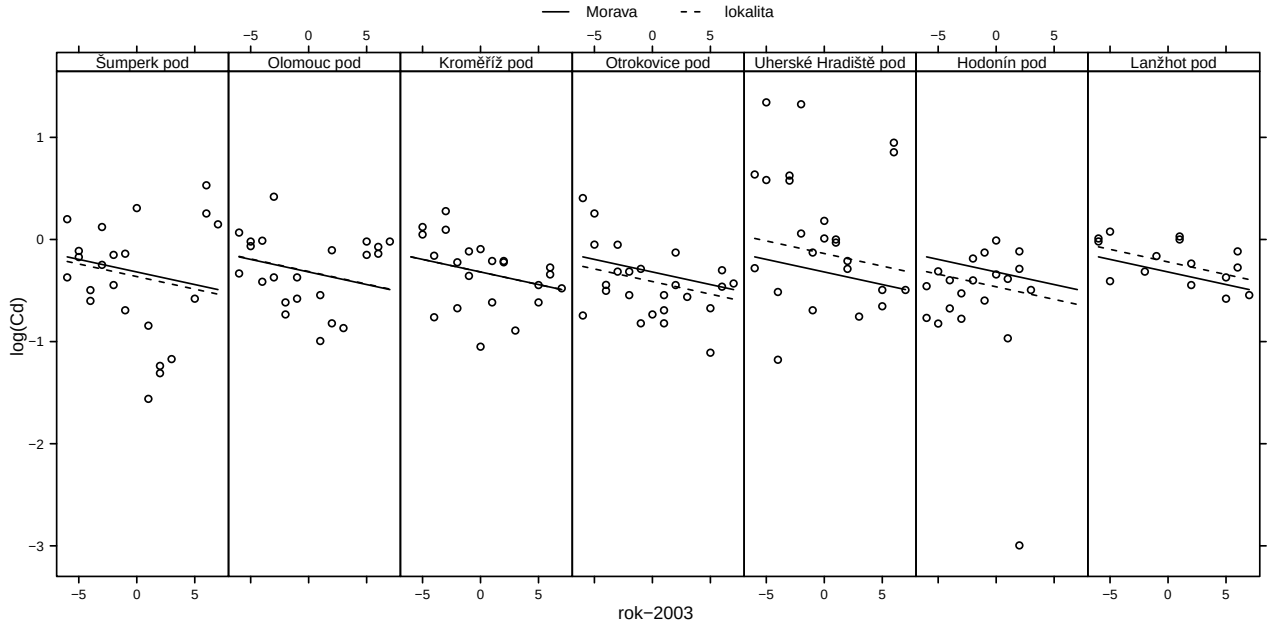
Lokalita	b_0 pro Pb	b_0 pro Cd	b_0 pro Hg	b_0 pro Ni
Šumperk pod	0.03621	−0.04501	−0.23950	−0.00006
Olomouc pod	0.04712	0.00467	0.11234	0.00002
Kroměříž pod	0.09957	−0.00008	0.18667	0.00008
Otrokovice pod	−0.02015	−0.09364	−0.04320	0.00011
Uherské Hradiště pod	0.09552	0.18078	0.09730	−0.00002
Hodonín pod	−0.13787	−0.14509	−0.23848	−0.00029
Lanžhot pod	−0.12040	0.09838	0.12488	0.00017

Znaménka predikovaných náhodných posunutí určují, zda predikovaná přímka j-té lokality $(\beta_0 + b_{j0}) + \beta_1 t$ leží nad či pod přímkou $\beta_0 + \beta_1 t$, která popisuje trend celé řeky Moravy.

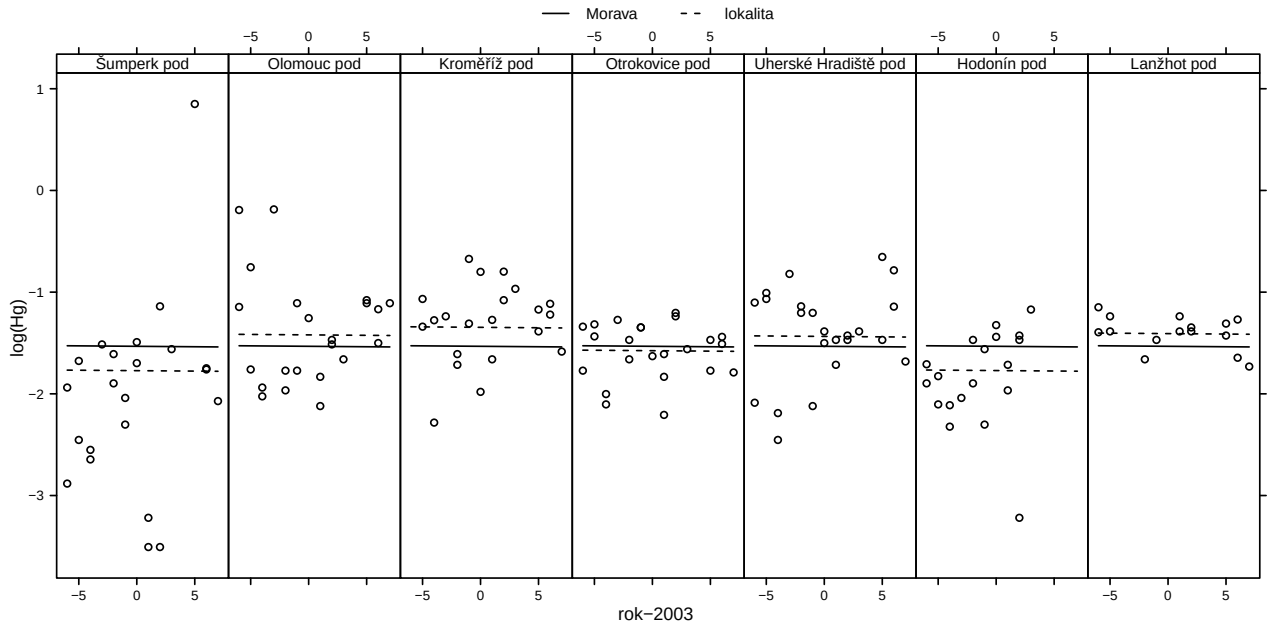
Na závěr ještě uvedeme čtyři grafy, které kromě naměřených hodnot graficky znázorňují odhady lokálních přímek $(\beta_0 + b_{j0}) + \beta_1 t$ (čárkovaná čára) a přímky $\beta_0 + \beta_1 t$ tvořené pevnými efekty (plná čára).



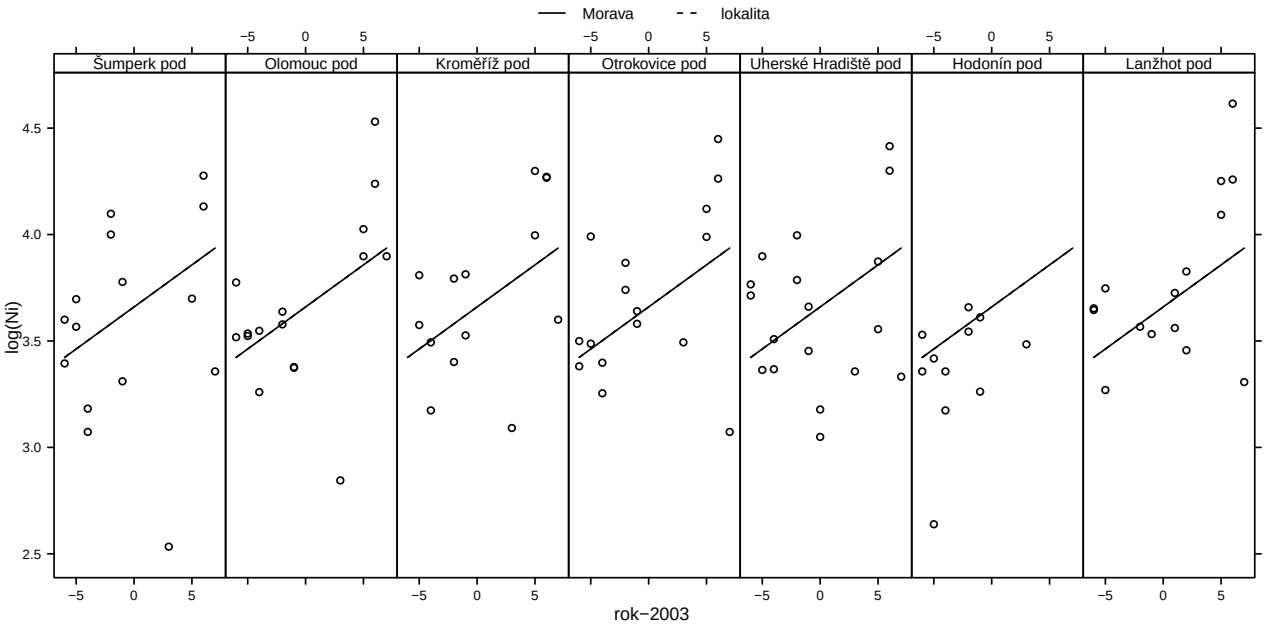
Obrázek 2: Odhady lineárního trendu logaritmu obsahu olova (Pb) v sedimentech v letech 1997–2010: společného celé řece Moravě (plná čára) i jednotlivých lokalit (čárkovaná čára) na základě LMM modelu.



Obrázek 3: Odhady lineárního trendu logaritmu obsahu kadmia (Cd) v sedimentech v letech 1997–2010: společného celé řece Moravě (plná čára) i jednotlivých lokalit (čárkovaná čára) na základě LMM modelu.



Obrázek 4: Odhady lineárního trendu logaritmu obsahu rtuti (Hg) v sedimentech v letech 1997–2010: společného celé řece Moravě (plná čára) i jednotlivých lokalit (čárkovaná čára) na základě LMM modelu.



Obrázek 5: Odhady lineárního trendu logaritmu obsahu niklu (Ni) v sedimentech v letech 1997–2010: společného celé řece Moravě (plná čára) i jednotlivých lokalit (čárkovaná čára) na základě LMM modelu.

Z posledního grafu je patrné, že u niklu je variabilita náhodných posunutí tak malá, že přímka $\beta_0+\beta_1 t$ (plná čára) vždy překrývá přímku $(\beta_0+b_{j0})+\beta_1 t$ (čárkovaná čára).

4. Závěr

V příspěvku je prezentováno využití lineárních smíšených modelů při statistické analýze dlouhodobých trendů zatížení sedimentů řeky Moravy prioritními a dalšími nebezpečnými látkami v letech 1997–2010, tedy v situaci, kdy jsou k dispozici korelovaná skupinová data s heteroskedastickými rozptyly. Pilotní statistické hodnocení bylo zaměřeno na ukazatele ze skupiny těžké kovy, které prokázalo klesající trend kadmia, mírně rostoucí trend niklu a minimálně rostoucí trend olova v monitorovaném úseku řeky Moravy. Obsah rtuti v sedimentech řeky Moravy se ve sledovaném období řádově nezměnil. Zvolený lineární regresní model s náhodnými a pevnými efekty, který je modifikací ANCOVA modelu, se ukázal jako optimální a bude využit i v rámci dalšího hodnocení zatížení sedimentů řeky Moravy prioritními a dalšími nebezpečnými látkami.

Veškeré výpočty byly provedeny pomocí volně šiřitelného programovacího prostředí R (viz [9] a [8]), které je snadno dostupné na internetu.

Literatura

[1] Bernardová I.: *Projekt Morava II.* DÚ 03. Hodnocení jakosti vody. Průběžná zpráva. VÚV T.G.M., Brno, 1996, 1997, 1998, 1999.

[2] Bernardová I.: *Projekt Morava III.* DÚ 04. Hodnocení jakosti vody. Průběžná zpráva. VÚV T.G.M., Brno, 2000, 2001, 2002.

[3] Bernardová I.: *Projekt Morava IV.* DÚ 04. Hodnocení jakosti vody. Průběžná zpráva. VÚV T.G.M., Brno, 2003, 2004, 2005, 2006.

[4] Demidenko E.: *Mixed models: Theory and Applications.* In: Wiley Series in Probability and Statistics, Hoboken, New Jersey, 2004.

[5] Hudcová H., Bernardová I.: *Projekt Identifikace antropogenních tlaků na kvalitativní stav vod a vodních ekosystémů v oblastech povodí Moravy a Dyje.* DÚ 7. Identifikace dopadů antropogenních tlaků na povrchové vody a vodní ekosystémy. Závěrečná syntetická zpráva o řešení dílčího úkolu za období 2008–2010. VÚV TGM, v.v.i., Brno, 2010.

[6] McCulloch C. E., Sealre S. R.: *Generalized, Linear, and Mixed Models.* In: Wiley Series in Probability and Statistics, Wiley, New York, 2001.

[7] Morrell Ch. H.: Likelihood Ratio Testing of Variance Components in the Linear Mixed-Effects Model Using Restricted Maximum Likelihood. *Biometrics* **54**, 1560–1568, 1998.

[8] Pinheiro J., Bates D., DebRoy S., Sarkar D. and the R Development Core Team: *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-103, 2012.

[9] R Development Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, 2012. URL: `http://www.R-project.org/`.

[10] Směrnice Evropského parlamentu a Rady 2008/105/ES o normách environmentální kvality v oblasti vodní politiky a o změně směrnice EP a Rady 2000/60/ES.

[11] Stram D. O., Lee J. W.: Variance component testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177, 1994.

[12] Verbeke G., Molenberghs G.: *Linear mixed models for longitudinal data.* In: Springer Series in Statistics, Springer, New York, 2000.

[13] Zdařil J. a kol.: *Projekt Morava.* Závěrečná zpráva. VÚV T.G.M., Brno, 67 s. + příl., 1996.

CITLIVOST PEARSONOVA CHÍ KVADRÁT TESTU NA VOLBU TŘÍD

SENSITIVITY OF PEARSON'S CHI-SQUARE TEST TO THE CHOICE OF CLASSES

Vít Kubelka

Adresa: MFF UK v Praze, Sokolovská 83, 186 75, Praha 8

E-mail: KubelkaVit@seznam.cz

Abstrakt: Tento článek se se zabývá závislostí p-hodnoty χ^2 testu dobré shody Poissonova rozdělení na volbě tříd. U náhodných výběrů různých rozsahů simulovaných z Poissonova rozdělení jsou spočítány p-hodnoty pro všechny volby tříd, které dodržují určenou minimální teoretickou četnost. Většinou je požadována minimální teoretická četnost 5, ale chování p-hodnoty je vyšetřováno i při dodržování minimálních teoretických četností 1, 10 a 20. Neznámý parametr je odhadován modifikovanou metodou minimálního χ^2 i výběrovým průměrem. Výsledky simulací ukazují, že podmínka minimální teoretické četnosti jednotlivých tříd není dostatečná a test je nespolehlivý. Proto je přidána podmínka téměř stejných teoretických četností v jednotlivých třídách. Na dalších simulacích je ukázáno, že při dodržování tohoto pravidla se spolehlivost testu výrazně zlepší.

Klíčová slova: testy dobré shody, Pearsonův chí kvadrát test, statistika chí kvadrát, volba tříd.

Abstract: The χ^2 goodness-of-fit test depends on the choice of classes. It is the purpose of this paper to investigate differences between p-values depending on the choice of classes in the case of Poisson distribution. Samples from Poisson distribution of various sizes are simulated. For all choices of classes which satisfy condition of minium expected frequency p-value is calculated. The simulations are made for various minimum expected frequencies. The unknown parameter is estimated both by modified method of minimum χ^2 and by sample mean. The simulations show that the condition of minimum expected frequency of classes is not sufficient and the test is unreliable. Therefore, the rule of almost equal expected frequencies in all classes is added, and in further simulations a substantial improvement of the test is shown if this condition is satisfied.

Keywords: Goodness-of-fit test, Pearson's chi-square test, chi-square statistic, choice of classes.

1. Úvod

Pearsonův χ^2 test dobré shody je jedním z hlavních nástrojů k ověření, zda náhodný výběr pochází z určitého rozdělení. Jeho základem je rozložit na intervaly výběrový prostor vyšetřovaného rozdělení a v těchto intervalech porovnat očekávané a zjištěné četnosti realizací zkoumané náhodné veličiny. Místo intervalů se většinou používá označení třídy. Na jejich volbě závisí p-hodnota daného testu. Mohou být zvoleny libovolně, pokud je dodržena minimální teoretická četnost jednotlivých tříd. Názory na to, jaká minimální teoretická četnost tříd by se měla dodržovat, se v literatuře dosti liší. Společné mají to, že jsou to pouze doporučení, i když renomovaných statistiků, nepodložená rigorózními důkazy. Přehled těchto doporučení uvádí Kubelka [3].

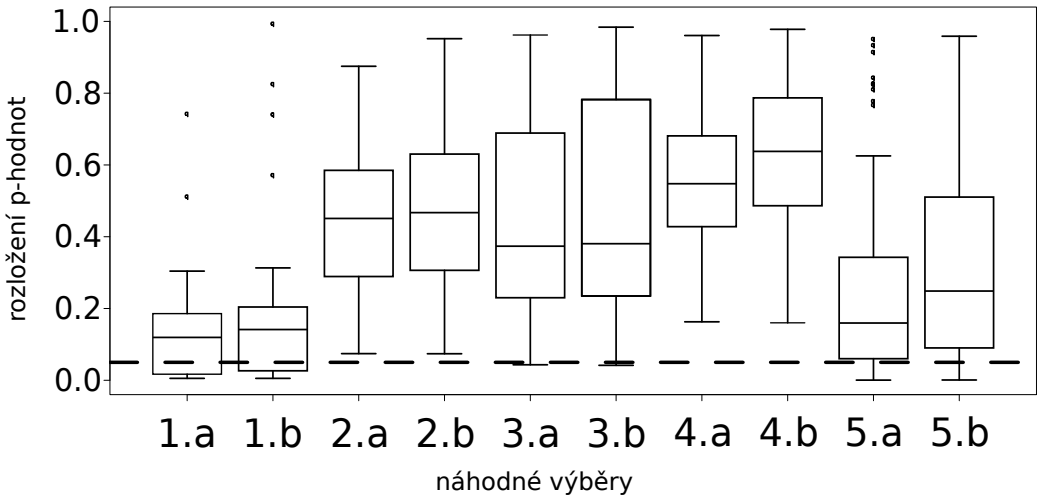
Důležitý je také správný odhad neznámých parametrů. V praxi se nezřídka používá odhad založený přímo na daném náhodném výběru (v případě Poissonova rozdělení je to například výběrový průměr), ale podle článku Chernoff a Lehmann [2] nemá χ^2 statistika v takovém případě asymptotické rozdělení, které se v tomto testu používá. Statistika χ^2 má toto asymptotické rozdělení, pokud jsou neznámé parametry odhadnuty na základě teoretických četností. Například, jak uvádí Anděl [1], pokud je odhad parametru proveden modifikovanou metodou minimálního χ^2, kterou budeme nadále značit MMMCH2.

Zaměříme se na χ^2 test dobré shody v případě Poissonova rozdělení. Mooney and Jolliffe [5] poukazují na velký rozdíl mezi získanými p-hodnotami pro dvě konkrétní volby tříd v případě náhodného výběru z praxe, u kterého zjišťují, zda se dá modelovat Poissonovým rozdělením. Avšak neznámý parametr odhadují právě výběrovým průměrem. Kubelka [3] pro stejný náhodný výběr počítá p-hodnoty pro všechny přípustné volby tříd. Neznámý parametr odhaduje výběrovým průměrem i MMMCH2 a ukazuje, že v závislosti na volbě tříd jsou rozdíly mezi získanými p-hodnotami velmi velké.

2. Citlivost testu na rozsah náhodného výběru

Budeme testovat náhodné výběry, u kterých víme, že pocházejí z Poissonova rozdělení s parametrem 7. Tomuto parametru je přibližně roven výběrový průměr náhodného výběru z článku Mooney and Jolliffe [5]. Neznámý parametr budeme odhadovat výběrovým průměrem i MMMCH2 a budeme dodržovat minimální teoretickou četnost 5.

Obrázek 1 vykresluje rozložení p-hodnot v závislosti na volbě tříd pro pět různých náhodných výběrů o rozsahu 50.



Obrázek 1: Rozložení p-hodnot v závislosti na volbě tříd, 5 různých náhodných výběrů o rozsahu 50 z Poissonova rozdělení s parametrem 7, minimální teoretická četnost 5. Písmeno a značí odhad parametru výběrovým průměrem, b značí MMMCH2. Vodorovná přerušovaná čára vyznačuje pětiprocentní hladinu významnosti testu.

Při použití obou metod odhadu parametru jsou rozdíly mezi jednotlivými p-hodnotami velmi výrazné. V případě prvního a posledního náhodného výběru výsledek testu zcela závisí na volbě tříd.

Za stejných podmínek byly testovány i pětice náhodných výběrů o rozsahu 25, 100, 500 a dvojice náhodných výběrů o rozsahu 5000. Ukázalo se, že rozsah náhodného výběru nemá velký vliv na rozložení p-hodnot v závislosti na volbě tříd a p-hodnoty se chovají podobně jako v případě náhodného výběru o rozsahu 50. Se zvětšujícím se rozsahem náhodného výběru se zvětšuje pouze rozdíl mezi odlehlými hodnotami. Porovnáme-li chování p-hodnot získaných při použití MMMCH2 a p-hodnot získaných pomocí parametru odhadnutého výběrovým průměrem; pro náhodné výběry o malém rozsahu jsou p-hodnoty získané pomocí výběrového průměru trochu méně závislé na volbě tříd, ale je o něco vyšší pravděpodobnost chyby prvního druhu. S rostoucím rozsahem náhodných výběrů tyto rozdíly mizí a pro náhodné výběry o rozsahu 500 již oba odhady dávají srovnatelné výsledky.

O volbách tříd, na kterých je nabýváno extremálních p-hodnot, nelze, dle našeho pozorování, udělat konkrétní závěr.

3. Vliv minimální požadované teoretické četnosti tříd

Stejným způsobem jako v předchozím odstavci byla programem testována trojice náhodných výběrů z Poissonova rozdělení s parametrem 7 o roz-

sahu 100, ale při dodržování různých teoretických četností. U každého z těchto třech náhodných výběrů byly porovnávány výsledky při dodržování minimální teoretické četnosti 1, 5, 10 a 20.

Ukázalo se, že kritérium minimální teoretické četnosti má na chování p-hodnot a na výsledek testu velmi malý vliv. Rozložení p-hodnot kolem svého výběrového průměru se nemění. Se zmenšující se požadovanou minimální teoretickou četností se zvětšuje pouze rozdíl mezi odlehlými pozorováními v rámci rozložení p-hodnot. V případě odhadu parametru výběrovým průměrem se lehce zvyšuje výběrový průměr p-hodnot, pokud požadujeme menší minimální teoretickou četnost. Tím se snižuje pravděpodobnost chyby prvního druhu. Avšak, protože pracujeme s náhodnými výběry, které z Poissonova rozdělení pocházejí, nemáme žádnou informaci o síle testu (tedy pravděpodobnosti chyby druhého druhu).

Jiné minimální teoretické četnosti než 5 tedy také nijak výrazně nesnižují rozdíly mezi jednotlivými p-hodnotami. Inspirujme se tedy článkem Mann and Wald [4] a zkusme navrhnout další pravidlo.

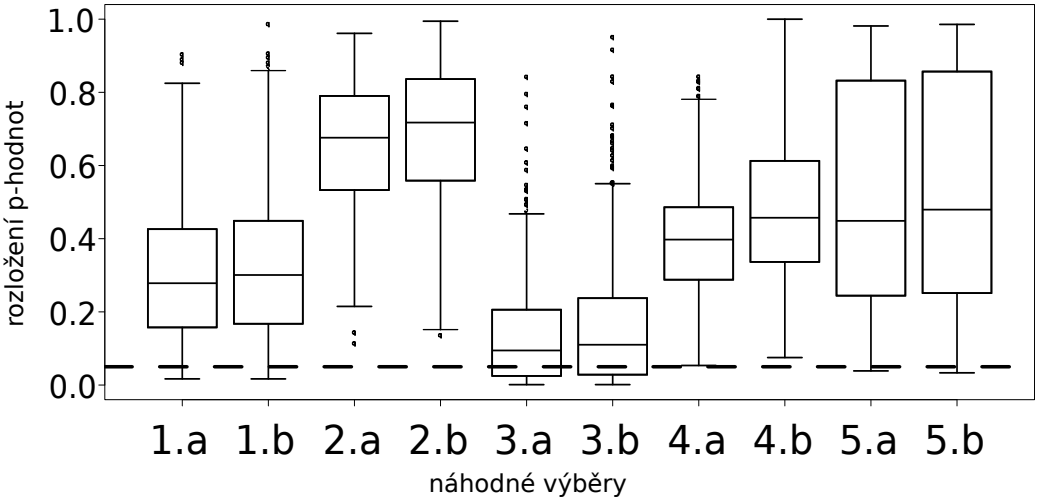
4. Rovnoměrné rozložení teoretických četností tříd

Mějme náhodný výběr o rozsahu n. Pro každou volbu o k třídách splňující požadovanou minimální teoretickou četnost spočítáme konstantu $1/k$. Projdeme všechny třídy a pokud se pravděpodobnost p_i některé z tříd liší od $1/k$ o více než C, kde $C \in [0, 1]$ je konstanta, kterou si předem určíme, potom takovou volbu tříd označíme za nevyhovující a nebudeme z ní počítat p-hodnotu. Teoretické četnosti jednotlivých tříd se potom neliší o více než $2nC$.

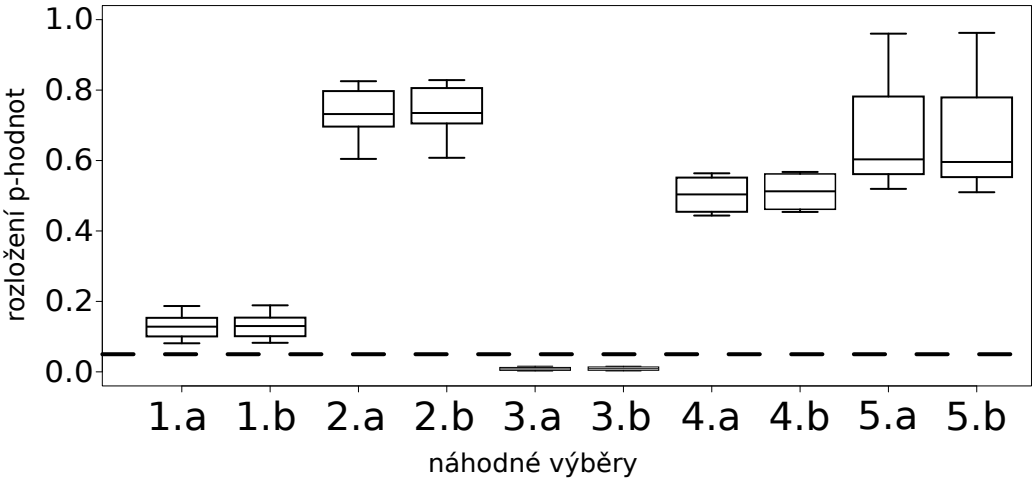
Vyšetříme pětici různých náhodných výběrů z Poissonova rozdělení o rozsahu 100. Obrázek 2 vykresluje rozložení p-hodnot v případě, kdy byla dodržena minimální teoretická četnost 5 a neznámý parametr byl odhadován výběrovým průměrem i MMMCH2, ale nebylo použito výše zmíněné pravidlo téměř stejných teoretických četností jednotlivých tříd.

Obrázek 3 vykresluje výsledky, pokud toto pravidlo aplikujeme. Konstantu C položíme rovnou 0,2.

Po aplikování našeho pravidla se rozdíl mezi jednotlivými p-hodnotami výrazně snížil. Částečně je to dáno tím, že při dodržování tohoto pravidla je výrazně méně možností, jak volit třídy. Avšak podstatné je, že p-hodnoty se přibližně stabilizovaly kolem jejich předchozích výběrových průměrů, ty přibližně odpovídají mediánům na obrázku 2. Všimněme si třetího náhodného výběru. V tomto případě je výsledek testu nejprve zcela závislý na volbě tříd. Avšak po aplikování pravidla téměř stejných teoretických četností se všechny



Obrázek 2: Rozložení p-hodnot v závislosti na volbě tříd bez použití pravidla téměř stejných teoretických četností, 5 různých náhodných výběrů o rozsahu 50 z Poissonova rozdělení s parametrem 7, minimální teoretická četnost 5. Písmeno a značí odhad parametru výběrovým průměrem, b značí MMMCH2. Vodorovná přerušovaná čára vyznačuje pětiprocentní hladinu významnosti testu.



Obrázek 3: Rozložení p-hodnot v závislosti na volbě tříd při dodržování pravidla téměř stejných teoretických četností, $C = 0{,}2$, pět různých náhodných výběrů o rozsahu 50 z Poissonova rozdělení s parametrem 7, minimální teoretická četnost 5. Písmeno a značí odhad parametru výběrovým průměrem, b značí MMMCH2. Vodorovná přerušovaná čára vyznačuje pětiprocentní hladinu významnosti testu.

p-hodnoty pohybují mezi 0,002 a 0,016 a pro všechny volby tříd tak zamítáme nulovou hypotézu (poznamenejme, že se v tomto případě dopouštíme chyby 1. druhu). Nicméně po vykreslení histogramu z tohoto náhodného výběru se ukázalo, že data skutečně výrazně neodpovídají Poissonovu rozdělení, přestože z něho byla generována.

Pravidlo bylo testováno i u náhodných výběrů z Poissonova rozdělení o rozsahu 50. Byla dodržována minimální teoretická četnost 5 i 1. Výsledky se podobaly výsledkům na obrázku 3, což ukazuje, že na požadované minimální teoretické četnosti příliš nezáleží, naopak je důležité rozložit teoretické četnosti jednotlivých tříd co nejrovnoměrněji. V závislosti na rozsahu náhodného výběru je nutné vhodně volit konstantu C tak, aby bylo pravidlo splněno alespoň pro některé volby tříd.

5. Závěr

Pokud u χ^2 testu požadujeme jako jedinou omezující podmínku pouze minimální teoretickou četnost jednotlivých tříd, test je nespolehlivý.

Při použití MMMCH2 je menší pravděpodobnost chyby prvního druhu než při odhadnutí parametru výběrovým průměrem, ale test více závisí na volbě tříd. Se zvyšujícím se rozsahem náhodného výběru se tyto rozdíly zmenšují. U rozsahu 500 již mezi metodami není rozdíl.

Je-li parametr odhadnut MMMCH2, téměř nezáleží na požadované minimální teoretické četnosti.

Je-li parametr odhadnut výběrovým průměrem, potom je-li požadována menší minimální teoretická četnost jednotlivých tříd, lehce se snižuje pravděpodobnost chyby prvního druhu.

Mezi volbami tříd příslušnými k jednotlivým extremálním p-hodnotám není viditelná souvislost.

Pokud volíme třídy tak, aby se jejich teoretické četnosti lišily co nejméně, průměrná p-hodnota je taková, jakou bychom očekávali, a rozdíly v p-hodnotách jsou minimální. Test tedy téměř nezávisí na volbě tříd a dává dobré výsledky. Výsledky jsou srovnatelné pro oba způsoby odhadu parametru i pro minimální teoretické četnosti 1 a 5. Pro minimální teoretickou četnost 1 je dokonce trochu menší pravděpodobnost chyby prvního druhu. Toto kritérium je tedy mnohem důležitější než způsob odhadu parametru a požadovaná minimální teoretická četnost jednotlivých tříd.

Podrobnější výsledky lze nalézt v bakalářské práci Kubelka [3].

Poděkování

Rád bych poděkoval panu profesoru Jiřímu Andělovi za cenné rady a čas, který mi věnoval.

Literatura

[1] Anděl J.: *Matematická statistika.* SNTL, Praha, 1978.

[2] Chernoff H., Lehmann E. L.: The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *Annals of Mathematical Statistics* **25**, 579–586, 1954.

[3] Kubelka V.: Citlivost testů dobré shody na volbu tříd. *Bakalářská práce.* MFF UK, Praha, 2014.

[4] Mann H. B., Wald A.: On the choice of the number of class intervals in the application of the chi square test. *Annals of Mathematical Statistics* **13**, 306–317, 1942.

[5] Mooney J., Jolliffe I.: Sensitivity of χ^2 goodness-of-fit test to the choice of classes. *Teaching Statistics* **26**, 22–23, 2004.

POZVÁNKA NA STAKAN 2015
INVITATION TO STAKAN 2015

Martina Litschmannová

E-mail: martina.litschmannova@vsb.cz

Milé kolegyně, vážení kolegové,

dovolte, abychom Vás pozvali na česko-slovenskou konferenci STAKAN 2015 (STAtističtí KANtoři), která se uskuteční ve dnech 9. – 11. října 2015 v Beskydech, hotel Čarták, Soláň, viz http://www.hotelcartak.cz/.

Konference STAtističtí KANtoři je společnou akcí České statistické společnosti a Slovenskej štatistickej a demografickej spoločnosti. Tyto společnosti organizují česko-slovenské konference každý lichý rok a země se pravidelně střídají. V České republice se konference koná pod názvem STAKAN a na Slovensku pod názvem PRASTAN.

V letošním roce je tato akce pořádána ve spolupráci s Katedrou aplikované matematiky z Fakulty elektrotechniky a informatiky Vysoké školy báňské – Technické univerzity v Ostravě.

Zaměření konference:

- ▷ Výuka statistiky na středních a vysokých školách.
- ▷ Aplikace statistických metod.
- ▷ Nové směry ve vývoji statistických metod.

Bližší informace, včetně on-line přihlašovacího formuláře, pokynů pro zaplacení vložného, zaslání abstraktu a zaslání příspěvku najdete na

http://www.statspol.cz/STAKAN15/

Těšíme se na Vaši účast!

Organizační výbor konference STAKAN 2015

Obsah

Vědecké a odborné články

Pozvánky na akce

Informační bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo. Vydavatelem je Česká statistická společnost, IČ 00550795, adresa společnosti je Na padesátém 81, 100 82 Praha 10. Evidenční číslo registrace vedené Ministerstvem kultury ČR dle zákona č. 46/2000 Sb. je E 21214. Časopis je na Seznamu recenzovaných neimpaktovaných periodik vydávaných v ČR, více viz server `http://www.vyzkum.cz/`.

The Information Bulletin of the Czech Statistical Society is published quarterly.
The contributions in bulletin are published in English, 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.