Zobecněné lineární modely, nebo analýza kompozičních dat?

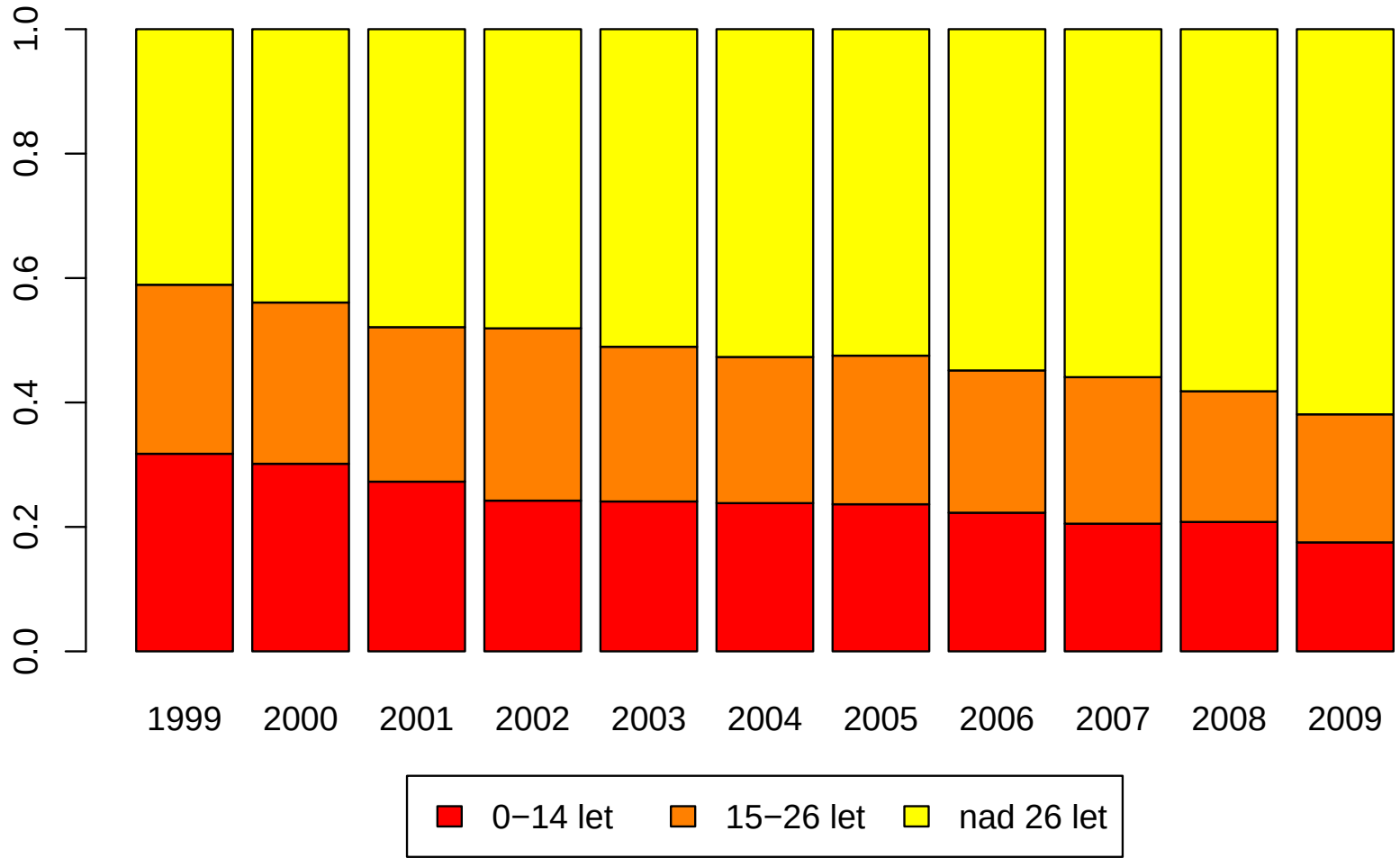
Podobnosti a rozdílnosti

Ondřej Vencálek
spolu s Karlem Hronem a Peterem Filzmoserem

Univerzita Palackého v Olomouci

Robust, Praděd, 12. září 2016

Hospitalizace cyklistů z důvodu poranění hlavy
vývoj zastoupení věkových skupin v letech 1999-2009



Zobecněný lineární model

- $\boldsymbol{Y}_i = (Y_{1,i}, Y_{2,i}, Y_{3,i})' \sim \mathrm{Multi}(n_i, \boldsymbol{\pi}_i), \quad i = 1, \ldots, N.$
- $E\,\boldsymbol{Y} = n\boldsymbol{\pi}$.
- Předpokládaný model:

$$\begin{aligned} \ln \frac{\pi_1(x)}{\pi_2(x)} &= \alpha_1 + \beta_1 x \\ \ln \frac{\pi_2(x)}{\pi_3(x)} &= \alpha_2 + \beta_2 x. \end{aligned}$$

- Adjacent-categories logit model („log-lineání model šancí sousedních kategorií“).
- Parametry $(\alpha_1, \alpha_2, \beta_1, \beta_2)$ odhadujeme metodou maximální věrohodnosti.

Analýza kompozičních dat

- $\boldsymbol{Y}_i = (Y_{1,i}, Y_{2,i}, Y_{3,i})'$, $\quad i = 1, \ldots, N$ považujeme za N pozorování 3-složkové kompozice.
- Kompozici lze vyjádřit pomocí různých „logratio“ souřadnic:
  - souřadnice založené na sousedních kategoriích:

$$Z_1(x) = \ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)}, \; Z_2(x) = \ln \frac{\widehat{\pi}_2(x)}{\widehat{\pi}_3(x)}$$

  - ortogonální souřadnice:

$$Z_1^*(x) = \ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)}, \; Z_2^*(x) = \ln \frac{\sqrt{\widehat{\pi}_1(x)\widehat{\pi}_2(x)}}{\widehat{\pi}_3(x)}$$

- $\widehat{\boldsymbol{\pi}}(x) = (\widehat{\pi}_1(x), \widehat{\pi}_2(x), \widehat{\pi}_3(x))' = \frac{1}{n(x)} \boldsymbol{Y}(x)$,
  tudíž $\frac{\widehat{\pi}_k(x_i)}{\widehat{\pi}_l(x_i)} = \frac{Y_{k,i}}{Y_{l,i}} \quad k, l = 1, \ldots, 3.$
  (předpokládáme nenulovost všech „složek“)

Analýza kompozičních dat (lineární model logratio souřad.)

$$Z_1(x_i) = \ln \frac{\widehat{\pi}_1(x_i)}{\widehat{\pi}_2(x_i)} = a_1 + b_1 x_i + \epsilon_{1,i},$$
$$Z_2(x_i) = \ln \frac{\widehat{\pi}_2(x_i)}{\widehat{\pi}_3(x_i)} = a_2 + b_2 x_i + \epsilon_{2,i}.$$

- $\boldsymbol{\epsilon}_i, i = 1, \ldots, N$ i.i.d.
- $\boldsymbol{\epsilon}_i = (\epsilon_{1,i}, \epsilon_{2,i})' \sim N(\mathbf{0}, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$

Tudíž:

$$\mathrm{E}Z_1(x) = \mathrm{E}\ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)} = a_1 + b_1 x$$
$$\mathrm{E}Z_2(x) = \mathrm{E}\ln \frac{\widehat{\pi}_2(x)}{\widehat{\pi}_3(x)} = a_2 + b_2 x.$$

Podobnost/rozdílnost

Sousední kategorie (zob.lin.model):

$$\ln \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1 + \beta_1 x$$
$$\ln \frac{\pi_2(x)}{\pi_3(x)} = \alpha_2 + \beta_2 x.$$

Sousední kategorie (kompoziční data):

$$\mathrm{E} \ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)} = a_1 + b_1 x$$
$$\mathrm{E} \ln \frac{\widehat{\pi}_2(x)}{\widehat{\pi}_3(x)} = a_2 + b_2 x.$$

Ortonormální souřadnice (zob.lin.model):

$$\ln \frac{\pi_1(x)}{\pi_2(x)} = \alpha_1^* + \beta_1^* x$$
$$\ln \frac{\sqrt{\pi_1(x)\pi_2(x)}}{\pi_3(x)} = \alpha_2^* + \beta_2^* x.$$

Ortonormální souřadnice (kompoziční data):

$$\mathrm{E} \ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)} = a_1^* + b_1^* x$$
$$\mathrm{E} \ln \frac{\sqrt{\widehat{\pi}_1(x)\widehat{\pi}_2(x)}}{\widehat{\pi}_3(x)} = a_2^* + b_2^* x.$$

Korespondence mezi parametry při použití různých logratio-souřadnic

$$\ln \frac{\pi_1(x)}{\pi_2(x)} = \ln \pi_1(x) - \ln \pi_2(x) = \alpha_1 + \beta_1 x$$

$$\ln \frac{\pi_2(x)}{\pi_3(x)} = \ln \pi_2(x) - \ln \pi_3(x) = \alpha_2 + \beta_2 x$$

$$\ln \frac{\pi_1(x)}{\pi_2(x)} = \ln \pi_1(x) - \ln \pi_2(x) = \alpha_1^* + \beta_1^* x$$

$$\ln \frac{\sqrt{\pi_1(x)\pi_2(x)}}{\pi_3(x)} = \frac{1}{2} \ln \pi_1(x) + \frac{1}{2} \ln \pi_2(x) - \ln \pi_3(x) = \alpha_2^* + \beta_2^* x$$

$$\alpha_1^* = \alpha_1,\ \beta_1^* = \beta_1,\ \alpha_2^* = \alpha_1/2 + \alpha_2,\ \beta_2^* = \beta_1/2 + \beta_2.$$

+ princip invariance pro maximálně věrohodné odhady:

$$\widehat{\alpha}_1^* = \widehat{\alpha}_1,\ \widehat{\beta}_1^* = \widehat{\beta}_1,\ \widehat{\alpha}_2^* = \widehat{\alpha}_1/2 + \widehat{\alpha}_2,\ \widehat{\beta}_2^* = \widehat{\beta}_1/2 + \widehat{\beta}_2.$$

Podobnost obou přístupů

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$
- $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\pi}_3)' = \boldsymbol{Y}/n$

$\boldsymbol{g} : \mathbb{R}^3 \to \mathbb{R}^2$

- Zob.lin.model: $\boldsymbol{g}(\boldsymbol{\pi}) = \left(\ln \frac{\pi_1}{\pi_2}, \ln \frac{\pi_2}{\pi_3}\right)'$
- An.kompoz.dat: $\boldsymbol{g}(\widehat{\boldsymbol{\pi}}) = \left(\ln \frac{\widehat{\pi}_1}{\widehat{\pi}_2}, \ln \frac{\widehat{\pi}_2}{\widehat{\pi}_3}\right)'$

Podobnost obou přístupů

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$
- $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\pi}_3)' = \boldsymbol{Y}/n$

$\sqrt{n}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_0)$,

kde $\boldsymbol{\Sigma}_0 = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$.

$\boldsymbol{g} : \mathbb{R}^3 \to \mathbb{R}^2$

- Zob.lin.model: $\boldsymbol{g}(\boldsymbol{\pi}) = \left(\ln \frac{\pi_1}{\pi_2}, \ln \frac{\pi_2}{\pi_3}\right)'$
- An.kompoz.dat: $\boldsymbol{g}(\widehat{\boldsymbol{\pi}}) = \left(\ln \frac{\widehat{\pi}_1}{\widehat{\pi}_2}, \ln \frac{\widehat{\pi}_2}{\widehat{\pi}_3}\right)'$

$\sqrt{n}(\boldsymbol{g}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{g}(\boldsymbol{\pi})) \xrightarrow{D} N(\mathbf{0}, (\partial \boldsymbol{g}/\partial \boldsymbol{\pi}) \boldsymbol{\Sigma}_0 (\partial \boldsymbol{g}/\partial \boldsymbol{\pi})')$,

kde $\partial \boldsymbol{g}/\partial \boldsymbol{\pi} = \begin{pmatrix} \frac{1}{\pi_1} & -\frac{1}{\pi_2} & 0 \\ 0 & \frac{1}{\pi_2} & -\frac{1}{\pi_3} \end{pmatrix}$

Podobnost obou přístupů

- $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)'$
- $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\pi}_3)' = \boldsymbol{Y}/n$

$\sqrt{n}(\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$,

kde $\boldsymbol{\Sigma}_0 = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$.

$\boldsymbol{g} : \mathbb{R}^3 \to \mathbb{R}^2$

- Zob.lin.model: $\boldsymbol{g}(\boldsymbol{\pi}) = \left(\ln \frac{\pi_1}{\pi_2}, \ln \frac{\pi_2}{\pi_3}\right)'$
- An.kompoz.dat: $\boldsymbol{g}(\widehat{\boldsymbol{\pi}}) = \left(\ln \frac{\widehat{\pi}_1}{\widehat{\pi}_2}, \ln \frac{\widehat{\pi}_2}{\widehat{\pi}_3}\right)'$

$\sqrt{n}(\boldsymbol{g}(\widehat{\boldsymbol{\pi}}) - \boldsymbol{g}(\boldsymbol{\pi})) \xrightarrow{D} N(\boldsymbol{0}, (\partial\boldsymbol{g}/\partial\boldsymbol{\pi})\boldsymbol{\Sigma}_0(\partial\boldsymbol{g}/\partial\boldsymbol{\pi})')$,

kde $\partial\boldsymbol{g}/\partial\boldsymbol{\pi} = \begin{pmatrix} \frac{1}{\pi_1} & -\frac{1}{\pi_2} & 0 \\ 0 & \frac{1}{\pi_2} & -\frac{1}{\pi_3} \end{pmatrix}$

Rozdílnost obou přístupů

$$\sqrt{n}\left(\begin{pmatrix} \ln\frac{\widehat{\pi}_1}{\widehat{\pi}_2} \\ \ln\frac{\widehat{\pi}_2}{\widehat{\pi}_3} \end{pmatrix} - \begin{pmatrix} \ln\frac{\pi_1}{\pi_2} \\ \ln\frac{\pi_2}{\pi_3} \end{pmatrix}\right) \xrightarrow{D} N\left(\mathbf{0}, \mathbf{\Sigma}_G(\boldsymbol{\pi})\right).$$

$$\mathbf{\Sigma}_G(\boldsymbol{\pi}) = \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_2} & -\frac{1}{\pi_2} \\ -\frac{1}{\pi_2} & \frac{1}{\pi_2} + \frac{1}{\pi_3} \end{pmatrix}$$

Připomenutí, předpokládaná var. matice v lineárním modelu pro logratio souřadnice (analýza kompozičních dat):

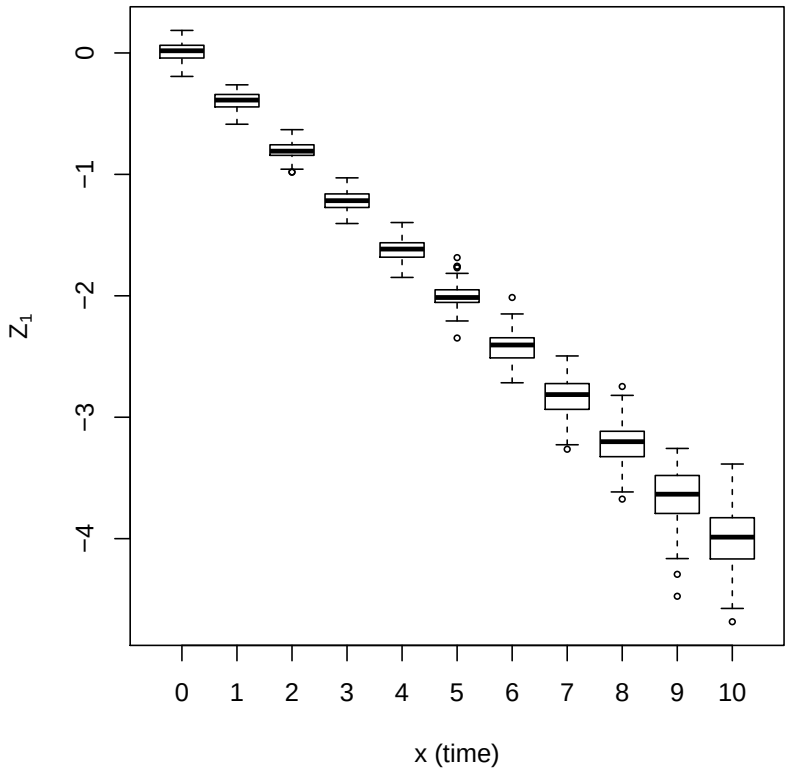
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

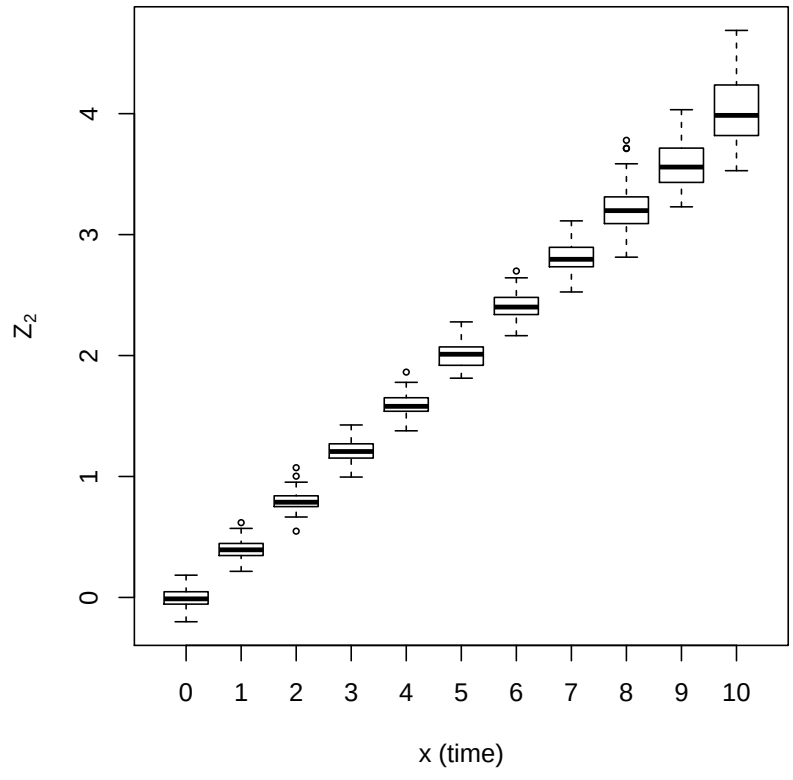
Numerická ilustrace I

$$\ln \frac{\pi_1(x)}{\pi_2(x)} = -0.4x, \qquad \ln \frac{\pi_2(x)}{\pi_3(x)} = \quad 0.4x.$$

x	π_1	π_2	π_3	$n\sigma_1^2$	$n\sigma_2^2$	ρ
0	0.33	0.33	0.33	6.00	6.00	-0.50
1	0.29	0.43	0.29	5.83	5.83	-0.60
2	0.24	0.53	0.24	6.12	6.12	-0.69
3	0.19	0.62	0.19	6.92	6.92	-0.77
4	0.14	0.71	0.14	8.36	8.36	-0.83
5	0.11	0.79	0.11	10.66	10.66	-0.88
6	0.08	0.85	0.08	14.20	14.20	-0.92
7	0.05	0.89	0.05	19.57	19.57	-0.94
8	0.04	0.92	0.04	27.61	27.61	-0.96
9	0.03	0.95	0.03	39.65	39.65	-0.97
10	0.02	0.96	0.02	57.63	57.63	-0.98

Numerická ilustrace I

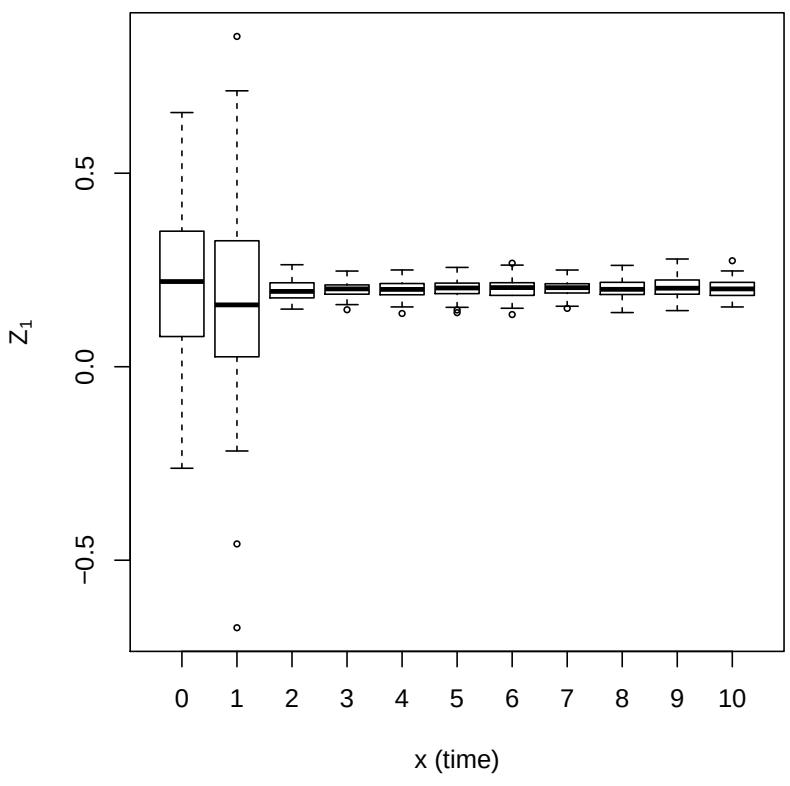


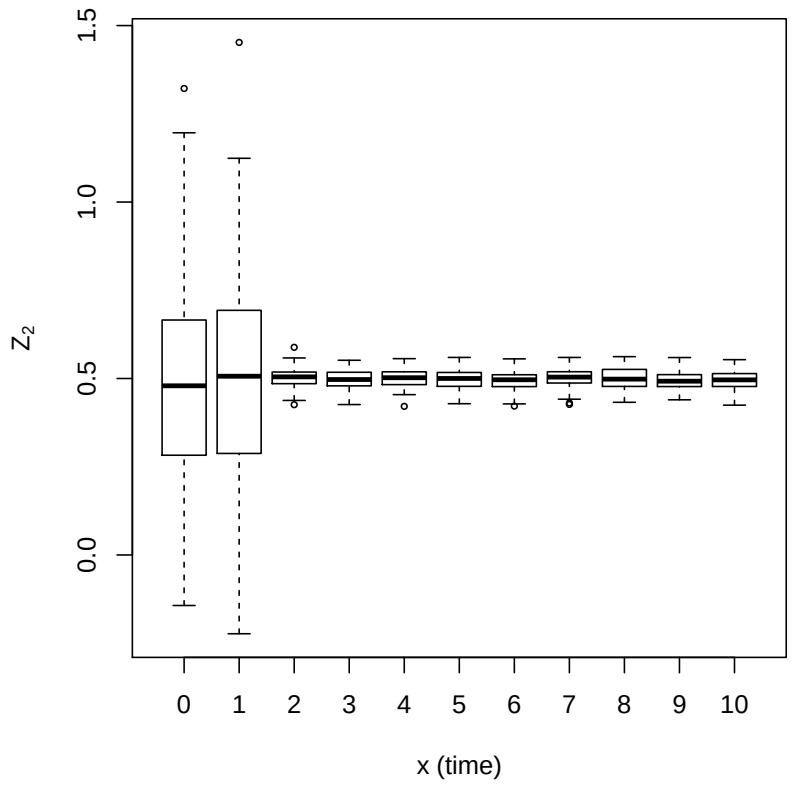


Numerická ilustrace II

- $\ln \frac{\pi_1(x)}{\pi_2(x)} = 0.2, \qquad \ln \frac{\pi_2(x)}{\pi_3(x)} = 0.5$
- $(\pi_1(x), \pi_2(x), \pi_3(x)) = (0.432, 0.354, 0.214)$ pro všechna x
- počet pozorování $n = 100$ pro $x \in \{0, 1\}$,
  $n = 10\,000$ pro $x \in \{2, \ldots, 10\}$
- při testování hypotézy nulovosti směrnice v lineárním modelu pro logratio souřadnice $H_1 : b_1 = 0$ získáváme 19-20 % (nesprávných) zamítnutí.
- stejný výsledek též pro test $H_2 : b_2 = 0$
- „důvod“...(viz následující obrázek)

Numerická ilustrace II





Hospitalizace cyklistů

Zobecněný lineární model:

$$\begin{aligned} \ln \frac{\pi_1(x)}{\pi_2(x)} &= \alpha_1 + \beta_1 x \\ \ln \frac{\pi_2(x)}{\pi_3(x)} &= \alpha_2 + \beta_2 x. \end{aligned}$$

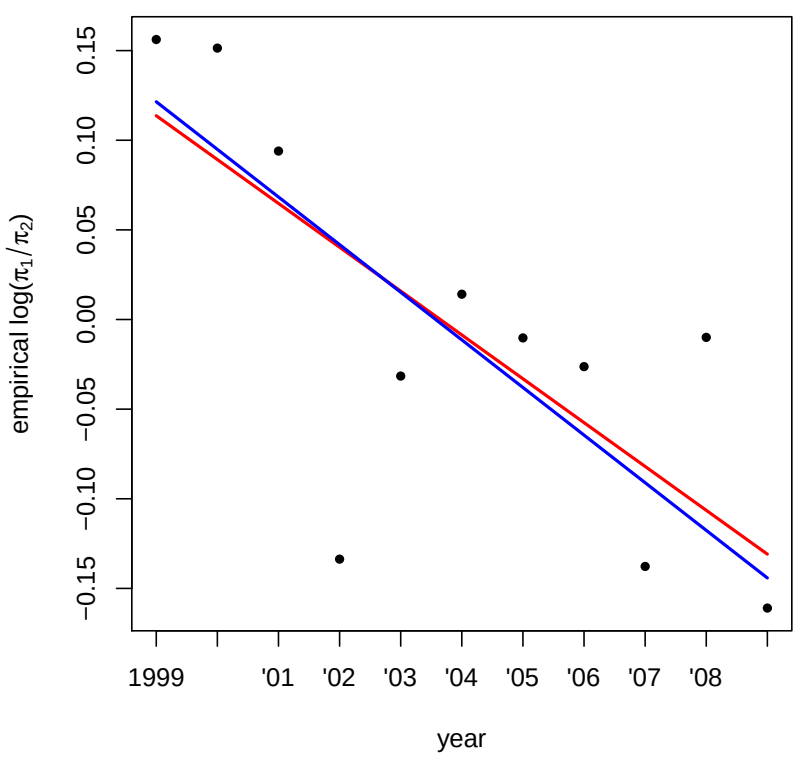
Lineární model pro logratio souřadnice (an. kompozičních dat):

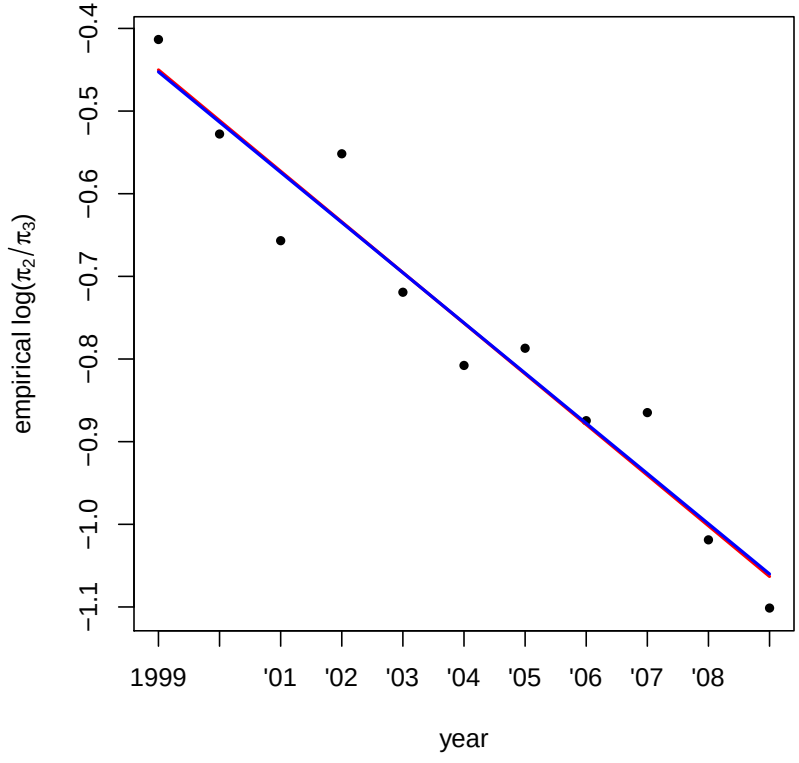
$$\begin{aligned} \mathrm{E} \ln \frac{\widehat{\pi}_1(x)}{\widehat{\pi}_2(x)} &= a_1 + b_1 x \\ \mathrm{E} \ln \frac{\widehat{\pi}_2(x)}{\widehat{\pi}_3(x)} &= a_2 + b_2 x. \end{aligned}$$

Odhady parametrů:

	kompoziční přístup	zob. lin. model	rozdíl
α_1 (a_1)	0.1137	0.1215	-0.0078
α_2 (a_2)	-0.4502	-0.4526	0.0023
β_1 (b_1)	-0.0245	-0.0266	0.0021
β_2 (b_2)	-0.0613	-0.0608	-0.0005

Hospitalizace cyklistů – porovnání modelů





modrá ... zobecněný lin. model

červená ... lin. model pro logratio souřadnice (an. kompozičních dat)

Hospitalizace cyklistů – porovnání modelů

Srovnání směrodatných odchylek pro odhady parametrů

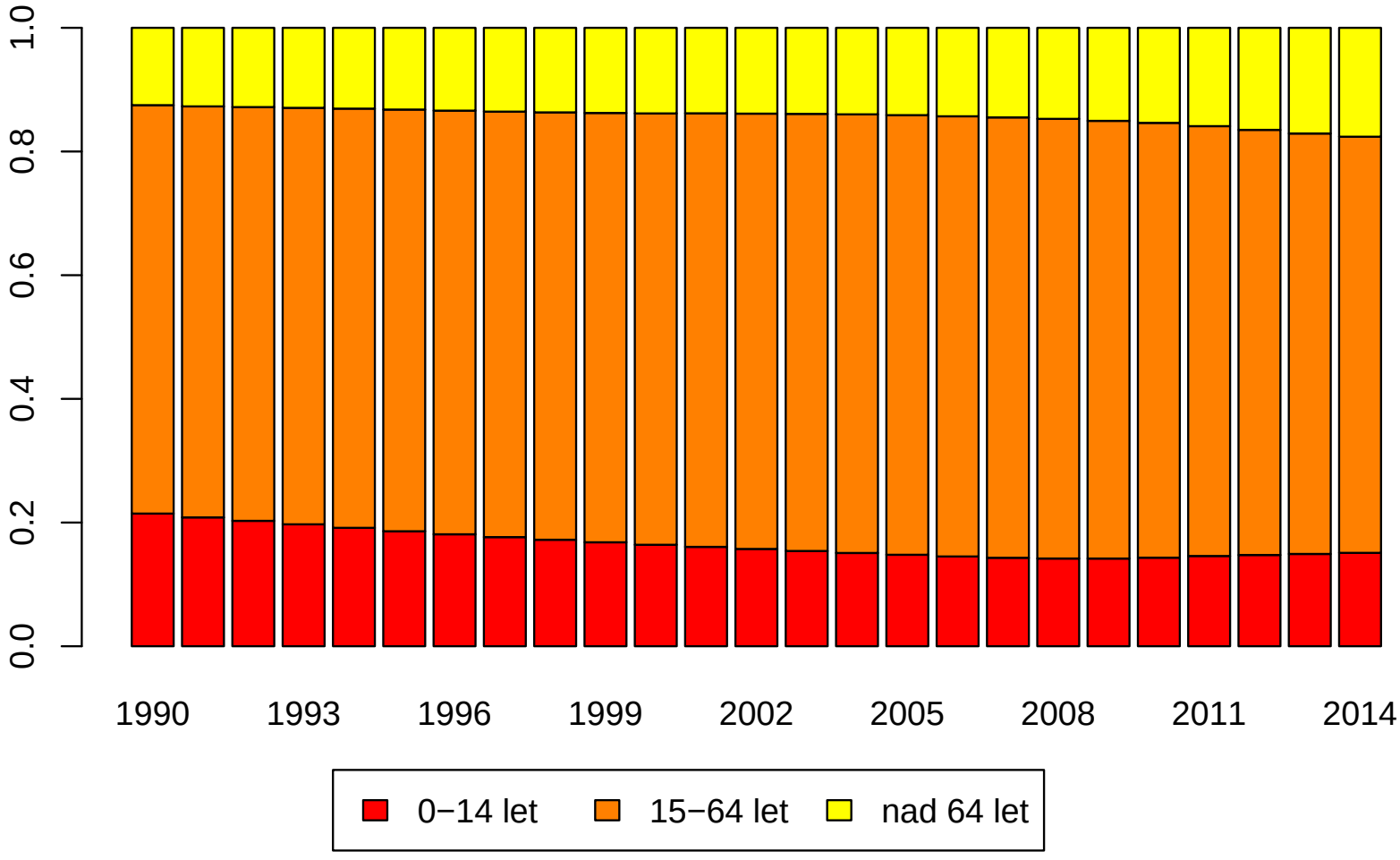
	kompoziční přístup	zob. lin. model	podíl
α_1 (a_1)	0.0438	0.0254	1.73
α_2 (a_2)	0.0310	0.0228	1.36
β_1 (b_1)	0.0074	0.0050	1.48
β_2 (b_2)	0.0052	0.0043	1.22

Další praktický příklad

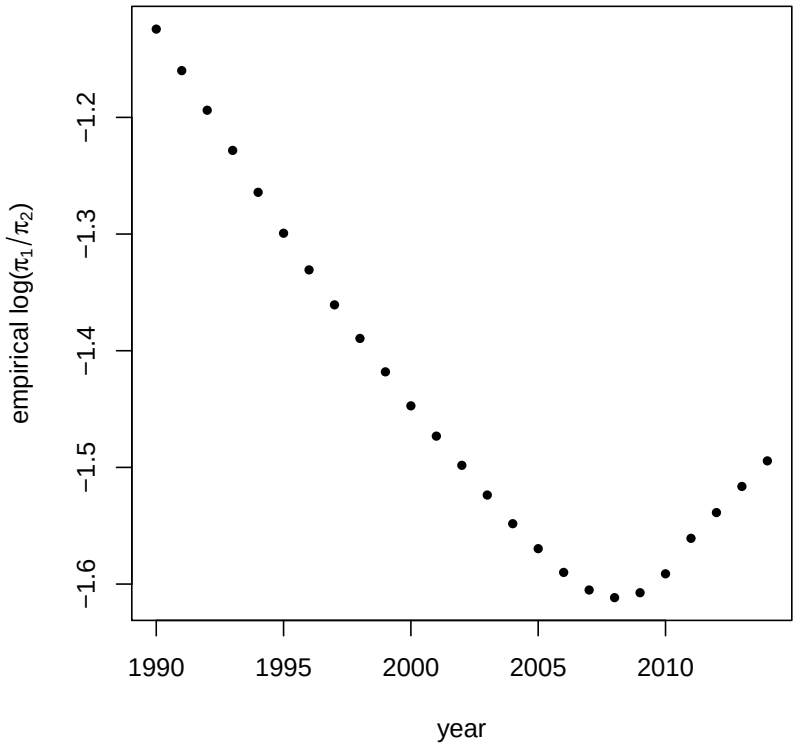
– **vývoj věkové sktruktury ČR od roku 1990 (do roku 2014)**

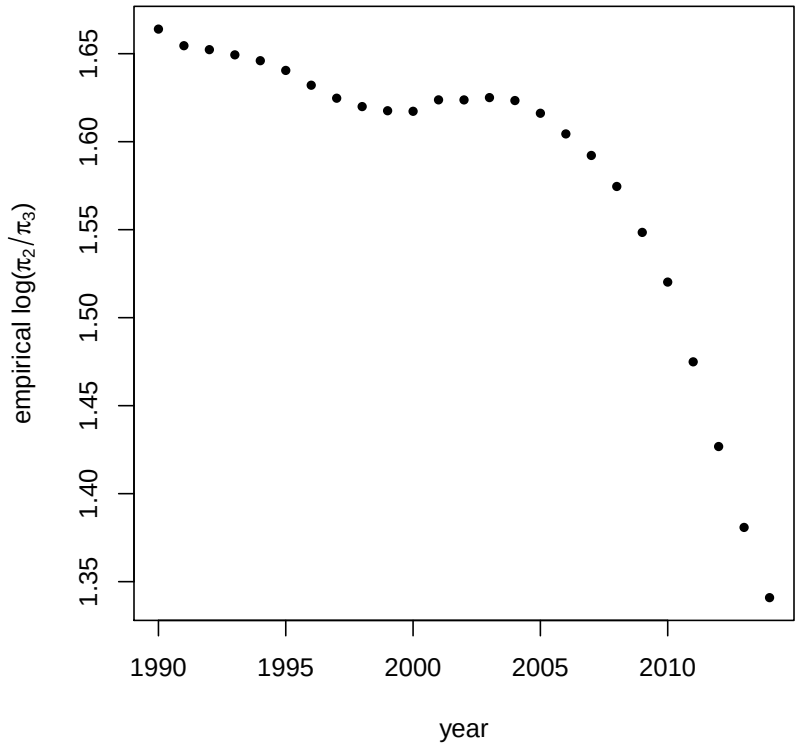
podle dat veřejně přístupných na stránkách ČSÚ

Vývoj věkové sktuktury ČR v letech 1990-2014



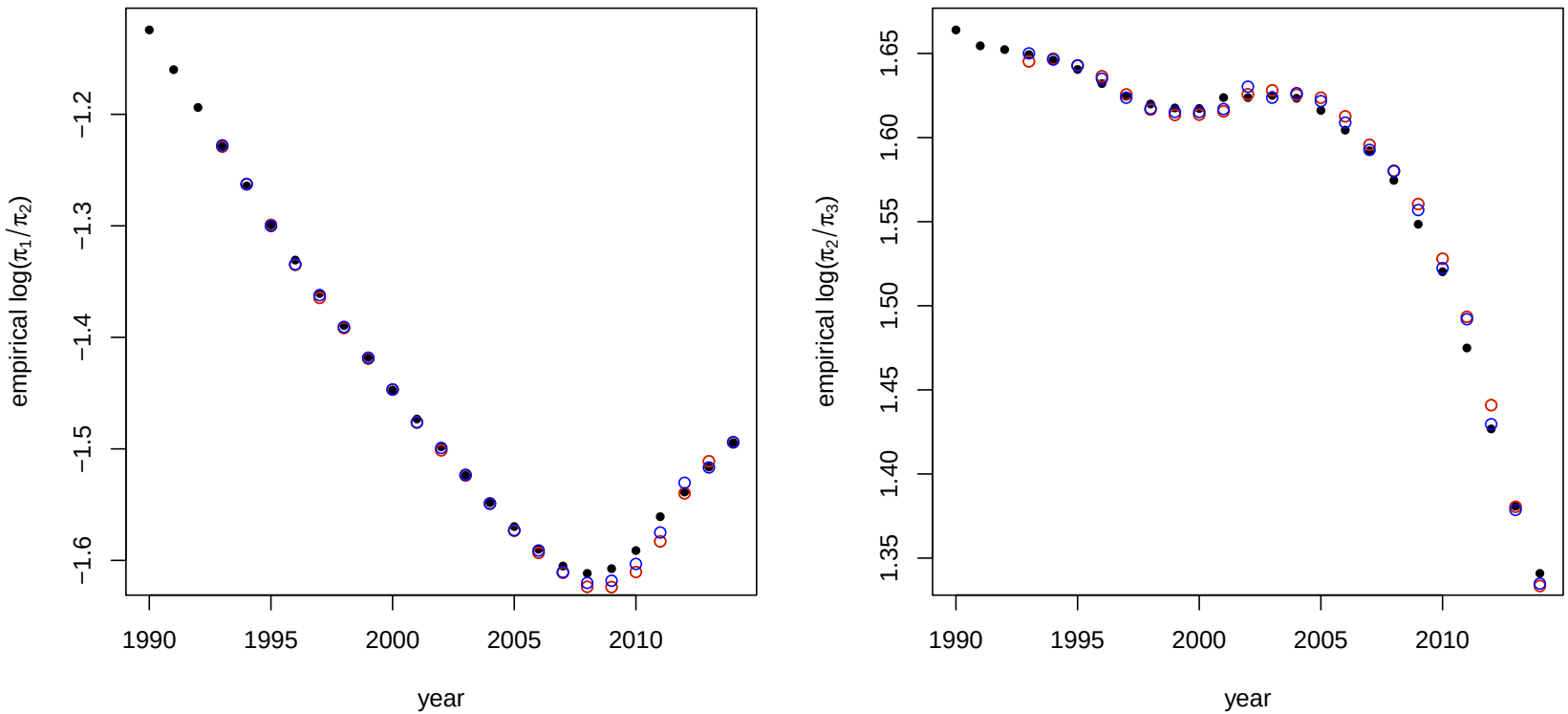
Vývoj věkové sktuktury ČR v letech 1990-2014, logratio souřadnice





Statistik versus opičák

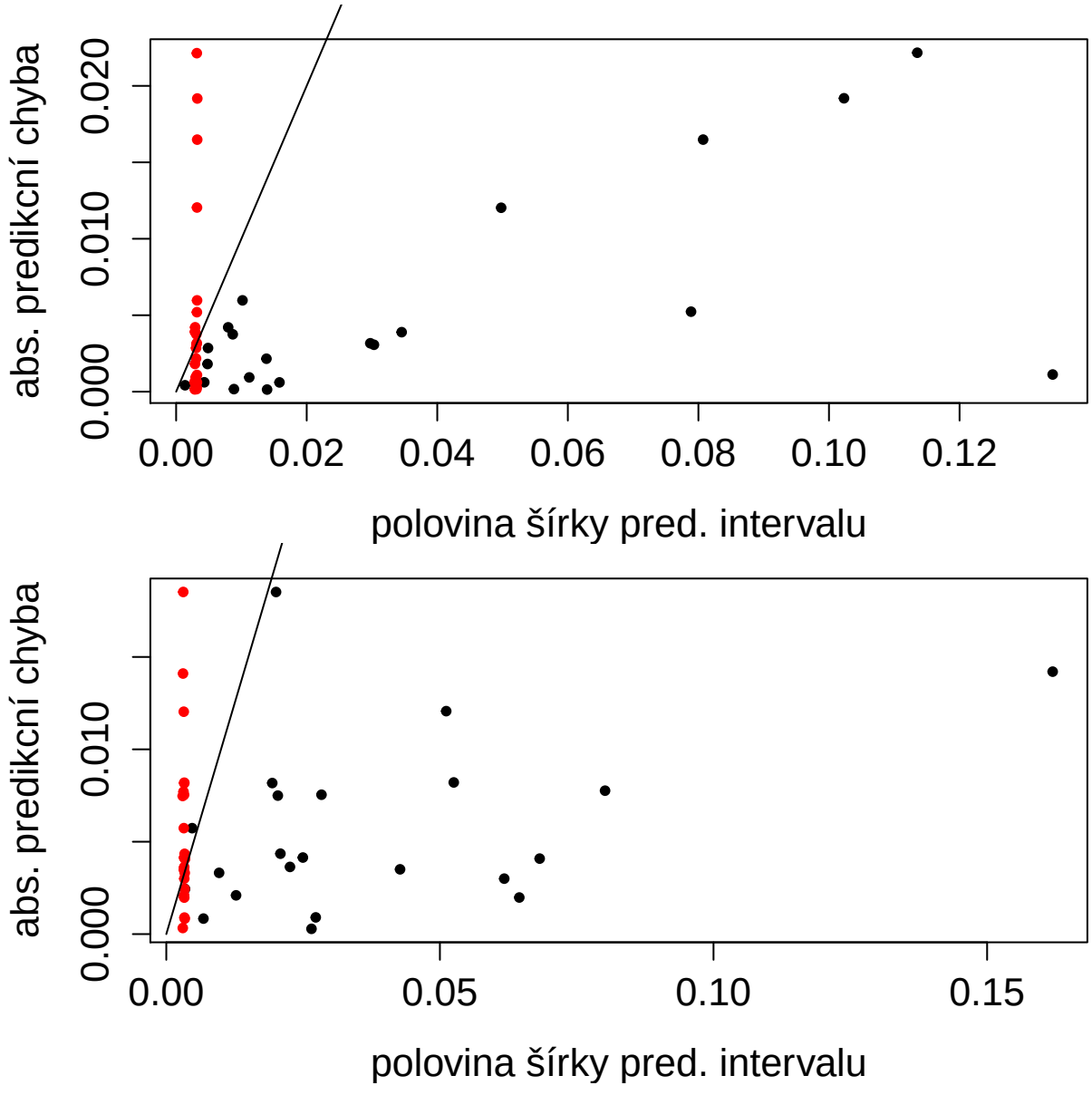
- **Statistik** má k dipozici sofistikované modely: zobecněný lineární model, analýzu kompozičních dat, používá lokálně lineární trend (červené body)
- **Opičák** předpovídá proložením posledních dvou hodnot přímkou (modré body)



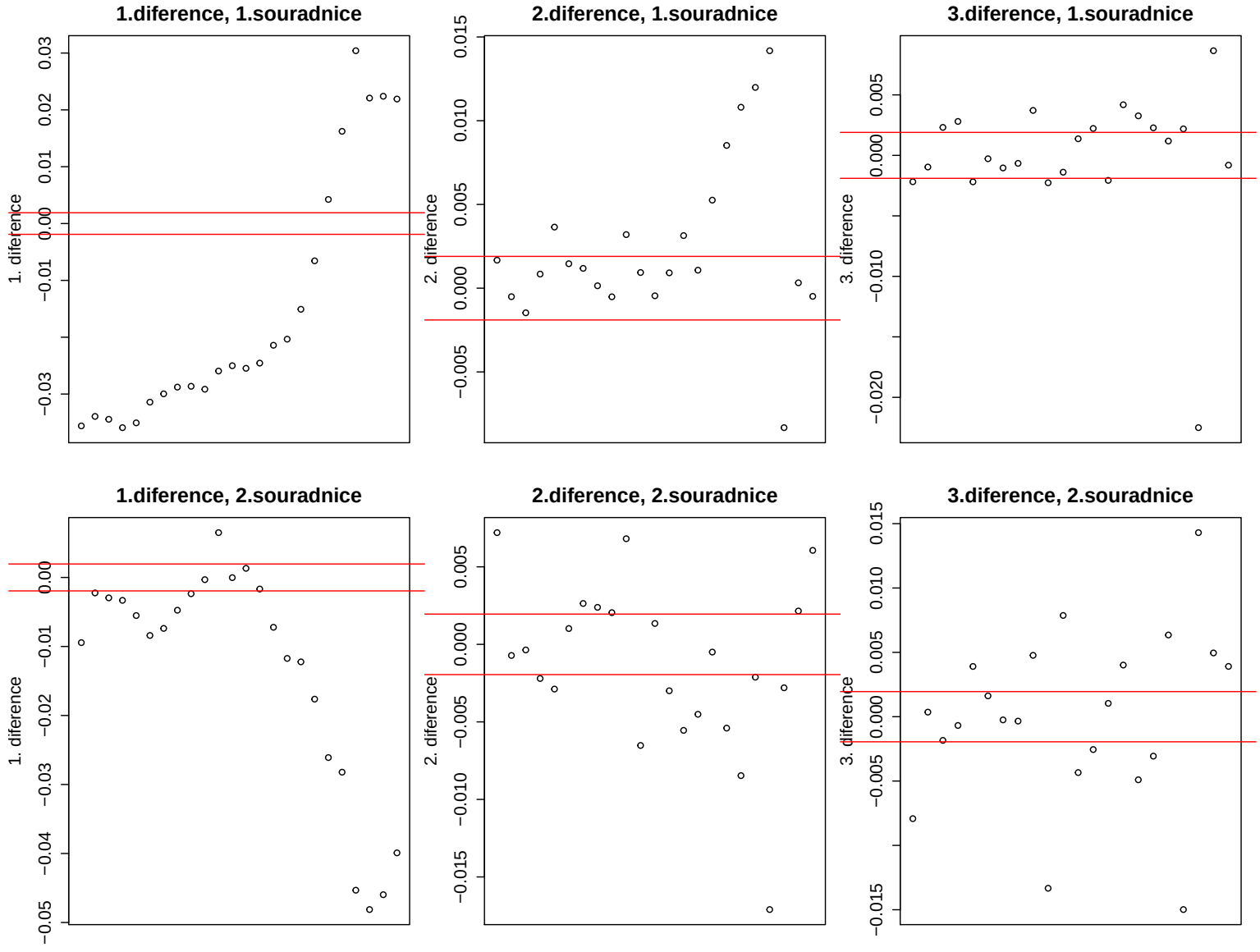
podíl průměrných absolutních predikčních chyb statistika proti opičákovi je 1,39 pro první souřadnici a 1,44 pro druhou souřadnici

rok	délka pred. intervalu		podíl délek
	kompoziční přístup	zob. lin. model	
1993	0.0317	0.0056	5.6
1994	0.0096	0.0058	1.7
1995	0.0279	0.0057	4.9
1996	0.0159	0.0058	2.7
1997	0.0691	0.0057	12.0
1998	0.0277	0.0061	4.5
1999	0.0224	0.0060	3.7
2000	0.0027	0.0061	0.4
2001	0.0098	0.0061	1.6
2002	0.0606	0.0062	9.7
2003	0.0177	0.0062	2.8
2004	0.0086	0.0064	1.4
2005	0.0173	0.0062	2.8
2006	0.0595	0.0063	9.4
2007	0.0203	0.0064	3.2
2008	0.0996	0.0064	15.6
2009	0.1614	0.0065	25.0
2010	0.2046	0.0065	31.6
2011	0.2270	0.0063	35.8
2012	0.2685	0.0064	42.1
2013	0.1577	0.0064	24.8
2014	0.0061	0.0063	0.96

(Abs.predikční chyby) versus (délky pred. intervalů)



Variabilita v datech



Zobecněné lineární modely, nebo analýza kompozičních dat? (závěr)

- Podbnost obou přístupů: užívají stejné logratio transformace.
- Modely se však přesto liší.
- Rozdíl je ve varianční struktuře var(transformace($\widehat{\boldsymbol{\pi}(x)}$)), která je v přístupu zobecněných lineárních modelů závislá na regresorech x a na počtu pozorování n, zatímco při analýze kompozičních dat ji pokládáme za nezávislou na x a n.
- Oba modely dávají v námi řešených praktických úlohách velmi podobné odhady parametrů a tedy i velmi podobné predikce, nedá se říci, že by odhady jednou z metod byly přesnější (ve smyslu nižší variability odhadů).
- Při hodně vysokých počtech pozorování (úloha týkající se věkové struktury populace ČR) může být přesnost predikce na základě zobecněného lineárního modelu nižší, než bychom dle modelu očekávali. Zdá se, že druhý přístup tímto problémem netrpí.