VZNIK LINEÁRNÍHO SPOJITÉHO TRENDU

Daniela JARUŠKOVÁ

FSt ČVUT, KM

Abstrakt. V článku se studují testy pro nalezení změny v jednoduché lineární regresi za předpokladu spojité regresní funkce, kde čas hraje roli nezávisle proměnné. Problém byl motivován snahou nalézt změny v meteorologických měřeních.

Abstract: Tests for detection of a change in simple linear regression are studied assuming the regression function is continuous at the change-point and the independent variable is equally spaced. Problem was motivated by effort of meteorologists to discover a change in meteorological measurements.

Резюме: В этой статье предлагаются тесты для нахождения разладки в модели линейной регрессии, которая непрерывна в точке разладки. Проблема выходит из анализа метеорологических данных.

1. Úvod.

Uvažujeme problém testování posloupnosti nezávislých stejně rozdělených náhodných veličin proti alternativě, že v neznámém čase dojde ke spojité změně ve střední hodnotě takové, že po okamžiku změny bude střední hodnota růst lineárně s časem. Přesněji řečeno to znamená, že chceme testovat nulovou hypotézu H_0 proti alternativě A:

$$\begin{aligned} H_0 &: X_i = \mu + e_i, && i = 1, ..., n, \\ A \quad &: \quad \exists\, k \in \{0, ..., n-1\} && \text{takové, že} \\ & X_i = \mu + e_i, && i = 1, ..., k, \\ & X_i = \mu + b \cdot (i - k) + e_i, && i = k+1, ..., n, \end{aligned} \tag{1.1}$$

kde e_i, $i = 1, ..., n$ jsou nezávislé stejně rozdělené, přičemž $\mathrm{E}\, e_i = 0, \mathrm{E}\, e_i^2 = \sigma^2$ a $\mathrm{E}\, |e_i|^{2+\delta} < \infty\, (\delta > 0)$. Alternativa může být jednostranná, jestliže $b > 0$ (resp. $b < 0$), nebo oboustranná, jestliže $b \neq 0$.

2. Případ známého μ.

Nejprve studujme případ, kde parametry μ a σ^2 jsou známé. Bez újmy na obecnosti můžeme předpokládat, že $\mu = 0$. Kdybychom uvažovali situaci, kde čas změny k je v modelu (1.1) známý, pak statistika

$$\hat{b}_k = \frac{\sum_{i=k+1}^{n} X_i(i-k)}{\frac{(n-k)(n-k+1)(2n-2k+1)}{6}} \tag{2.1}$$

je odhadem parametru b metodou nejmenších čtverců. V případě, že čas změny nezná-me, pak je přirozené použít pro testování nulové hypotézy proti jednostranné alternativě $b > 0$ statistiku

$$\max_{0 \le k \ \le n-1} \tilde{b}_k / \sigma, \tag{2.2}$$

kde

$$\tilde{b}_k = \frac{\sum_{i=k+1}^{n} X_i(i-k)}{\sqrt{\frac{(n-k)(n-k+1)(2n-2k+1)}{6}}} \tag{2.3}$$

Poznamenejme, že tato statistika je ekvivalentní s poměrem věrohodnosti, jestliže chyby e_i, $i = 1, ..., n$ jsou normálně rozdělené. Pro oboustrannou alternativu můžeme použít obdobně

$$\max_{0 \le k \ \le n-1} |\tilde{b}_k| / \sigma. \tag{2.4}$$

Zřejmě platí

$$\mathrm{E}\,\tilde{b}_k = 0,\ \mathrm{Var}\,\tilde{b}_k = \sigma^2,$$

$$\mathrm{Corr}\,(\tilde{b}_k, \tilde{b}_l) = \frac{\frac{(n-l)(n-l+1)(2(n-l)+1)}{6} + (l-k)\frac{(n-l)(n-l+1)}{2}}{\sqrt{\frac{(n-k)(n-k+1)(2(n-k)+1)}{6}}\sqrt{\frac{(n-l)(n-l+1)(2(n-l)+1)}{6}}},\ k \ \le l. \tag{2.5}$$

Rozdělení statistiky $\max_{0 \le k \ \le n-1} \tilde{b}_k / \sigma$ (resp. $\max_{0 \le k \ \le n-1} |\tilde{b}_k| / \sigma$) je velmi komplikované, a proto je přirozené studovat asymptotické rozdělení této statistiky pro velká n. Na intervalu $\langle 0, 1)$ definujme proces $\tilde{b}_n(t)$, $t \in \langle 0, 1)$ tak, že $\tilde{b}_n(t) = \tilde{b}_{[nt]}$. Tento proces pro každé $T < 1$ konverguje v distribuci na $D\langle 0, T\rangle$ ke standardizovanému gaussovskému procesu

$$IW(t) = \frac{\int_t^1 (W(1) - W(s))\, ds}{\sqrt{\frac{(1-t)^3}{3}}} = \frac{\int_t^1 (s\ - t)\, dW(s)}{\sqrt{\frac{(1-t)^3}{3}}}, \quad t \in \langle 0, 1). \tag{2.6}$$

($W(t), t \ge 0$ označuje Wienerův proces.) Platí $\overline{\lim}_{t \to 1} IW(t) = \infty$ s.j., stejně jako $\lim_{n \to \infty} \max_{0 \le k \ \le n-1} \tilde{b}_k / \sigma = \infty$ s.j. Abychom dostali nedegenerované limitní rozdělení, je třeba zkoumat překročení hranice, která s n roste nade všechny meze.

Platí

$$\lim_{n \to \infty} P\left(\max_{0 \le k \ \le n-1} \tilde{b}_k / \sigma > u_n \right) = 1 - \exp\left(-e^{-x}\right), \tag{2.7}$$

$$\lim_{n\to\infty} P\left(\max_{0\le k\ \le n-1} |\tilde{b}_k|/\sigma > u_n\right) = 1 - \exp\left(-2e^{-x}\right), \tag{2.8}$$

kde

$$u_n = \sqrt{2\ln\ln n} + \frac{1}{\sqrt{2\ln\ln n}}\left(\ln\frac{\sqrt{3}}{4\pi} + x\right), \quad x \in R_1. \tag{2.9}$$

Pokud parametr σ neznáme, je možno jej nahradit odhadem

$$\hat{\sigma}^2 = \sqrt{\sum X_i^2/n}$$

nebo uvažovat statistiku

$$\max_{0\le k\ \le n-1} \tilde{b}_k/\hat{\sigma}_k$$

kde $\hat{\sigma}_k = \sqrt{(\sum X_i^2 - \tilde{b}_k^2)/(n-1)}$. Limitní vztahy (2.7) a (2.8) zůstanou v platnosti.

Jiná možnost se nabízí v případě, kdy víme jistě, že ke změně nemohlo dojít v posledních $(1-\alpha)100\%$ časových okamžikách. Pak je vhodné uvažovat statistiku $\max_{0\le k\ \le[(1-\alpha)n]} \tilde{b}_k/\sigma$. Kritické hodnoty je možno odvodit pomocí aproximace limitním procesem

$$P\left(\max_{0\le k\ \le[(1-\alpha)n]} \tilde{b}_k/\sigma > x\right) = P\left(\max_{0\le t\le(1-\alpha)} IW(t) > x\right), \tag{2.10}$$

přičemž

$$P\left(\max_{0\le t\le(1-\alpha)} IW(t) > x\right) \simeq \sqrt{\frac{3}{8}}\frac{\phi(x)}{\sqrt{\pi}}\ln\frac{1}{\alpha}, \tag{2.11}$$

kde $\phi(x)$ je hustota $N(0,1)$ rozdělení.

3. PŘÍPAD NEZNÁMÉHO μ.

Jestliže parametr μ neznáme, je odhadem b metodou nejmenších čtverců v modelu (1.1)

$$\hat{B}_k = \frac{\sum_{i=k+1}^{n}(X_i - \bar{X})(i-k)}{\frac{(n-k)(n-k+1)(2n-2k+1)}{6} - \frac{(n-k)^2(n-k+1)^2}{4n}}. \tag{3.1}$$

Přirozenou testovou statistikou pro jednostrannou alternativu $b > 0$ je

$$\max_{0\le k\ \le n-1} \tilde{B}_k/\sigma, \tag{3.2}$$

resp. pro oboustrannou alternativu $b \neq 0$,

$$\max_{0\le k\ \le n-1} |\tilde{B}_k|/\sigma, \tag{3.3}$$

kde

$$\tilde{B}_k = \frac{\sum_{i=k+1}^{n}(X_i - \bar{X})(i-k)}{\sqrt{\frac{(n-k)(n-k+1)(2n-2k+1)}{6} - \frac{(n-k)^2(n-k+1)^2}{4n}}}. \tag{3.4}$$

Opět budeme zkoumat rozdělení těchto statistik pro velká n. Na intervalu $\langle 0,1)$ definujeme proces $\tilde{B}_n(t)$, $t \in \langle 0,1)$ tak, že $\tilde{B}_n(t) = \tilde{B}_{[nt]}$. Pro každé $T < 1$ proces $\tilde{B}_n(t)$ konverguje v distribuci na $D\langle 0,T\rangle$ ke standardizovanému gaussovskému procesu

$$IB(t) = \frac{\int_t^1 (s-t)dW(s) - W(1)\frac{(1-t)^2}{2}}{\sqrt{\frac{(1-t)^3}{3} - \frac{(1-t)^4}{4}}}, \quad t \in \langle 0,1). \tag{3.5}$$

Vzhledem k tomu, že opět platí $\lim_{n\to\infty} \max_{0\le k\ \le n-1} \tilde{B}_k/\sigma = \infty$ s.j., je třeba pro získání nedegenerovaného rozdělení uvažovat pravděpodobnosti meze překročení, která jde s rostoucím n do nekonečna. Platí

$$\lim_{n\to\infty} P\left(\max_{0\le k\ \le n-1} \tilde{B}_k/\sigma > u_n\right) = 1 - \exp\left(-e^{-x}\right), \tag{3.6}$$

$$\lim_{n\to\infty} P\left(\max_{0\le k\ \le n-1} |\tilde{B}_k|/\sigma > u_n\right) = 1 - \exp\left(-2e^{-x}\right), \tag{3.7}$$

kde

$$u_n = \sqrt{2\ln\ln n} + \frac{1}{\sqrt{2\ln\ln n}}\left(\ln\frac{\sqrt{3}}{4\pi} + x\right), \quad x \in R_1.$$

Pokud parametr σ neznáme, je možno jej nahradit odhadem

$$\hat{\sigma}^2 = \sqrt{\sum (X_i - \bar{X})^2/n}$$

nebo uvažovat statistiku

$$\max_{0\le k\ \le n-1} \tilde{B}_k/\hat{\sigma}_k,$$

kde

$$\hat{\sigma}_k = \sqrt{(\sum (X_i - \bar{X})^2 - \tilde{B}_k^2)/(n-2)}.$$

Limitní vztahy (3.6) a (3.7) zůstanou v platnosti.

Tabulka 1 porovnává kritické hodnoty statistiky $\max_{0\le k\ \le n-1} \tilde{B}_k/\hat{\sigma}_k$ pro některé hodnoty n spočtené z asymptotického rozdělení (3.7) s kritickými hodnotami získanými simulací. Simulované kritické hodnoty byly odhadnuty odpovídajícími kvantily empirické distribuční funkce statistiky

$$\max_{0\le k\ \le n-1} \tilde{B}_k/\hat{\sigma}_k$$

vytvořené na základě 100 000 realizací.

	5 % kritické hodnoty	
n	by (3.7)	simul.
100	2.71	2.63
200	2.75	2.65
300	2.77	2.65
500	2.79	2.68

	1 % kritické hodnoty	
n	by (3.7)	simul.
100	3.64	3.21
200	3.64	3.22
300	3.64	3.22
500	3.64	3.22

Tabulka 1. Několik příkladů 5 % and 1 % kritických hodnot statistiky $\max_{0\le k\ \le n-1} |\tilde{B}_k|/\widehat{\sigma}_k$ spočtených pomocí (3.7) a odhadnutých ze simulace.

V případě, kdy víme jistě, že ke změně nemohlo dojít v posledních $(1-\alpha)100\%$ časových okamžikách, můžeme uvažovat statistiku $\max_{0\le k\ \le[(1-\alpha)n]} \tilde{B}_k/\sigma$. Kritické hodnoty je možno odvodit pomocí aproximace limitním procesem

$$P\left(\max_{0\le k\ \le[(1-\alpha)n]} \tilde{B}_k/\sigma > x\right) = P\left(\max_{0\le t\le(1-\alpha)} IB(t) > x\right), \tag{3.8}$$

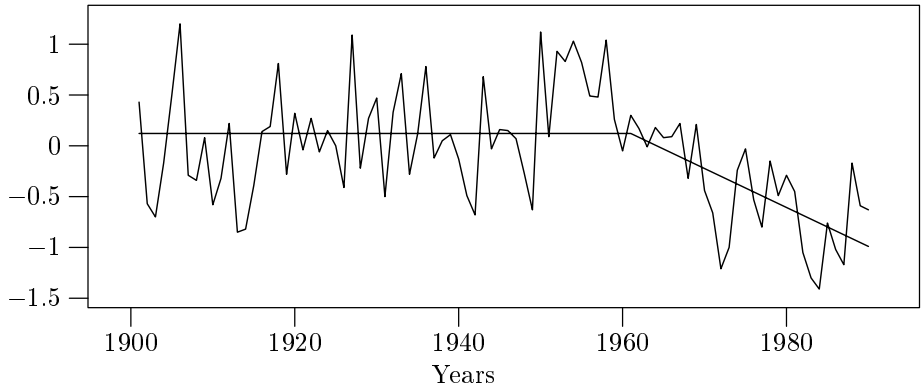
přičemž

$$P\left(\max_{0\le t\le(1-\alpha)} IB(t) > x\right) \simeq (1-\Phi(x)) +$$

$$+\left(\frac{\sqrt{3}}{2}\log\left(\frac{1+\sqrt{1-\alpha}}{1-\sqrt{1-\alpha}}\right) - \arctan\sqrt{3(1-\alpha)}\right)\frac{e^{-x^2/2}}{2\pi}. \tag{3.9}$$

4. PŘÍKLAD.

Příkladem postupné lineární změny může sloužit postupné zmenšování množství srážek v oblasti Sahelu. Obrázek 1 ukazuje vývoj standardizovaných ročních srážkových odchylek pro období 1901−1990 ($\frac{x_i-\bar{x}}{s}$, $i = 1, ...,$ 90) sestavený Nicolsonem (1993).



Obrázek 1. Standardizované roční srážkové odchylky v Sahelu, 1901 – 1990.

Statistická analýza prokazuje lineární úbytek srážek, přičemž změna nastala kolem roku 1960. Hodnota statistiky $\max_{0\le k\ \le n-1}|\tilde{B}_k|/\hat{\sigma}_k = 6.4452$ vysoko překračuje 1% kritickou hodnotu 3.64 získanou z (3.7). Hodnoty autoregresní funkce spočtené z residuí pro několik prvních zpoždění jsou: $ar(1) = 0.21, ar(2) = 0.13, ar(4) = 0.9, ar(5) = 0.13, ar(6) = -0.03$, $ar(7) = -0.01$ ... Zdá se tedy, že by původní veličiny měly být spíš modelovány nějakou ARMA posloupností. Vzhledem k tomu, že hodnoty autoregresní funkce nejsou příliš vysoké, závěr o statisticky významném snížení množství srážek zůstane v platnosti, viz Antoch et al (1995).

REFERENCE

Antoch, J., M. Hušková and Z. Prášková (1996) "Effect of dependence on statistics for determination of change", *Journal of Statistical Planning and Inference* (to appear).

Gombay, E. and L. Horváth (1994b), "Limit theorems for change in linear regression". *Journal of Multivariate Analysis* **48**, 43−69.

Davies R.B. (1987), "Hypothesis testing when a nuisance parameter is present only under the alternative." *Biometrika* **74**, 33−43.

Jarušková, D. (1996), "Some problems with application of change-point detection methods to environmental data". *Envirometrics* (to appear).

Kim, H.J. and D.Siegmund (1989), "The likelihood ratio test for a change-point in simple linear regression". *Biometrika* **76**, 409−423.

Nicholson, S.E. (1994),"Century-scale series of standardized annual departure of African rainfall". In *Trends'93: A Compendium of Data on Global Change,* T.A. Boden, D.P. Kaiser, R.J. Sepanski and F.W. Stoss (eds.) 952−962. ORNL/CDIAC - 65, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, Oak Ridge, Tenn., U.S.A.