STATISTICKÉ MODELY ZMĚN

Marie Hušková

Jarníkovská přednáška

10.10.2007

Outline

1. Úvod

Outline

Outline

1. Úvod
2. Změny v regresi, Ráztoka
3. Pravděpodobnostní nástroje

Outline

Outline

1. Úvod
2. Změny v regresi, Ráztoka
3. Pravděpodobnostní nástroje
4. Ráztoka - pokračování
5. Klementinská data

Outline

Outline

Úvod

- Change point problem
- Structural breaks
- Disorder problem
- Switching regression, segmented regression

Úvod

- Change point problem
- Structural breaks
- Disorder problem
- Switching regression, segmented regression

Typický případ:

$Y_1, \ldots, Y_n$ data získaná v časových okamžicích
$t_1 < t_2 < \ldots < t_n$

Úvod

Statistické modely změn

Outline
Úvod
Regrese
Pravděpodob nástroje
Raztoka
Klementinum
Finanční časové řady

- Change point problem
- Structural breaks
- Disorder problem
- Switching regression, segmented regression

Typický případ:

$Y_1, \ldots, Y_n$ data získaná v časových okamžicích
$t_1 < t_2 < \ldots < t_n$

- Jsou data generována jediným pravděpodobnostním mechanismem nebo je lze rozdělit do dvou skupin, t.j. existuje $k^*(< n)$ takové, že $Y_1, \ldots, Y_{k^*}$ a $Y_{k^*+1}, \ldots, Y_n$ jsou generovány dvěma různými pravděpodobnostními mechanismy?

Úvod

- Change point problem
- Structural breaks
- Disorder problem
- Switching regression, segmented regression

Typický případ:

$Y_1, \ldots, Y_n$ data získaná v časových okamžicích
$t_1 < t_2 < \ldots < t_n$

- Jsou data generována jediným pravděpodobnostním mechanismem nebo je lze rozdělit do dvou skupin, t.j. existuje $k^*(< n)$ takové, že $Y_1, \ldots, Y_{k^*}$ a $Y_{k^*+1}, \ldots, Y_n$ jsou generovány dvěma různými pravděpodobnostními mechanismy?

- Odhadnout k^* a další charakteristiky modelů.

Aplikace v mnoha oborech

- statistická kontrola jakosti;
- klimatologie, hydrologie, oblast životní prostředí;
- ekonometrie, modely finančních časových řad;
- medicínský výzkum a pod.

Počátky souvisí se statistickou kontrolou jakosti (1930)

Počátky souvisí se statistickou kontrolou jakosti (1930)

Wald (1940-1949) — teoretické základy sekvenční statistiky

Změny v regresi — Quandt (1958), Chow (1960), Hinkley (1969, 1971), Brown, Durbin, Evans (1975) -

Mc Neill (1974)— změny v polynomické regresi

Cobb (1978) – data o Nilu

Bagshaw, Johnson (1974)— závislá data

Počátky souvisí se statistickou kontrolou jakosti (1930)

Wald (1940-1949) — teoretické základy sekvenční statistiky

Změny v regresi — Quandt (1958), Chow (1960), Hinkley (1969, 1971), Brown, Durbin, Evans (1975) -

Mc Neill (1974)— změny v polynomické regresi

Cobb (1978) – data o Nilu

Bagshaw, Johnson (1974)— závislá data

Systematický vývoj od konce sedmdesátých let

Velký rozvoj od konce osmdesátých let — počítače a dostupné metody z teorie pravděpodobnosti

V současnosti nejvíce prací kolem modelování změn v modelech finančních řad, modelování změn kolem životního prostředí

Skupina v ČR: J. Antoch, D. Jarušková, Z. Prášková a jejich současní i bývalí studenti doktorského i magisterského studia a občas další tuzemští statistici

Skupina v ČR: J. Antoch, D. Jarušková, Z. Prášková a jejich současní i bývalí studenti doktorského i magisterského studia a občas další tuzemští statistici

Mezinárodní spolupráce: A. Aue (USA), I. Berkes (Rakousko), di Bucchianico (Nizozemí), E. Gombay (Kanada), L. Horváth (USA), C. Kirch (Německo), P. Kokoszka (USA), T. Ledwina (Polsko), S. Meintanis (Řecko), J. Steinebach (Německo), N. Veraverbeke (Belgie), W.R. van Zwet (Nizozemí)

Skupina v ČR: J. Antoch, D. Jarušková, Z. Prášková a jejich současní i bývalí studenti doktorského i magisterského studia a občas další tuzemští statistici

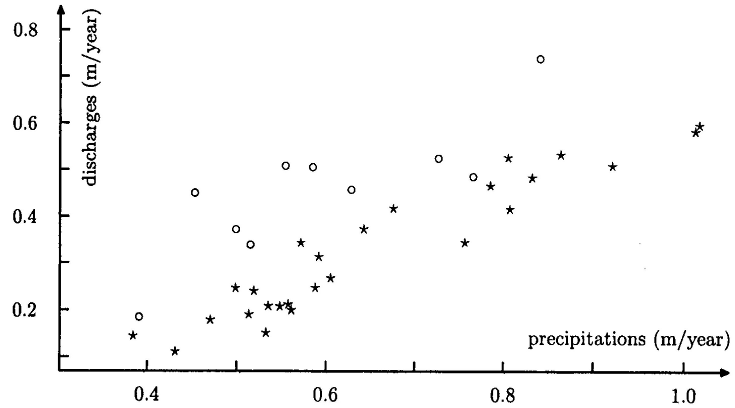
Mezinárodní spolupráce: A. Aue (USA), I. Berkes (Rakousko), di Bucchianico (Nizozemí), E. Gombay (Kanada), L. Horváth (USA), C. Kirch (Německo), P. Kokoszka (USA), T. Ledwina (Polsko), S. Meintanis (Řecko), J. Steinebach (Německo), N. Veraverbeke (Belgie), W.R. van Zwet (Nizozemí)

Granty GAČR, NATO a další mezinárodní granty

Outline

Potok Malá Ráztoka v Beskydech

Změny v regresních modelech, Ráztoka - data



Malá Ráztoka: Data.

Data z potoka Malá Ráztoka v letech 1954 — 1989 (36 let)
(x_i, Y_i), $i = 1, \ldots, 36$,
x_i — srážky za rok $\quad$ Y_i — odtok za rok

Došlo v průběhu 36 let ke změně v závislosti odtoku na srážkách?

$$
\begin{aligned}
E\,Y_i &= \beta_1 + \beta_2 x_i, & i &= 1, \ldots, k^* \\
E\,Y_i &= \beta_1 + \beta_2 x_i + \delta_1 + \delta_2 x_i, & i &= k^* + 1, \ldots, n
\end{aligned}
$$

$(\beta_1, \beta_2), (\delta_1, \delta_2), k^*$ – parametry

$H_0 : k^* = n \quad \& \quad H_1 : k^* < n$

odhad parametrů: $k^*, \beta_1, \beta_2, \delta_1, \delta_2$

Obecný lineární model se změnou

$Y_1, \ldots, Y_n$ pozorování v $t_1 < \cdots < t_n$:

$$
\begin{aligned}
Y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + e_i, & i = 1 \ldots, k^* \\
&= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\delta} + e_i, & i = k^* + 1 \ldots, n,
\end{aligned}
$$

$\mathbf{x}_i^T \boldsymbol{\beta}$ — skalární součin vektorů $\boldsymbol{x}_i$ a $\boldsymbol{\beta}_i$, T označuje transpozici

$e_1, \ldots, e_n$ – náhodné chyby splňující např. nezávislost a momentové podmínky (nulová střední hodnota, rozptyl σ^2 a $E|e_i|^{2+\delta} < \infty$ pro nějaké $\delta > 0$)

$\boldsymbol{\beta}, \boldsymbol{\delta} \neq \mathbf{0}$ — parametry

k^* — bod změny (change point)

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ — p-dim. regresní vektory (náhodné nebo nenáhodné):

- *regrese bez trendu*: $\frac{1}{n}\sum_{i=1}^{k} \mathbf{x}_i \mathbf{x}_i^T \approx \frac{k}{n}\mathbf{C}, \; k \leq n$

- *regrese s trendem*: $\mathbf{x}_i = \mathbf{h}(i/n), i = 1, \ldots, n$, $\mathbf{h}$ — vektor hladkých funkcí

- obvykle první složka x_i je 1

- Chceme testovat H_0 : žádná změna & H_1: aspoň jedna změna (testová statistika a kritická hodnota)

- odhad bodu změny k^* a ostatních parametrů

Testové statistiky obvykle založené na modifikovaném podílu věrohodností:

$$T_n = \max_{p \le k < n-p} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)^T \left(\widehat{\boldsymbol{\Sigma}}_k\right)^{-1} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)$$

$\widehat{\boldsymbol{\beta}}_k$ — odhad $\boldsymbol{\beta}$ založený na $Y_1, \ldots, Y_k$ (L_2 , L_1)
$\widehat{\boldsymbol{\beta}}_k^0$ — odhad $\boldsymbol{\beta}$ založený na $Y_{k+1}, \ldots, Y_n$ (L_2, L_1)
$\widehat{\boldsymbol{\Sigma}}_k$ — odhad varianční matice $\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0$

useknutá verze T_n:

$$T_n(\varepsilon) = \max_{n\varepsilon \le k \le n(1-\varepsilon)} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)^T \left(\widehat{\boldsymbol{\Sigma}}_k\right)^{-1} (\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)$$

$\varepsilon \in (0, 1/2)$

T_n blízko $T_n^0 = \max_{p \le k < n-p} \boldsymbol{S}_k^T \left(\widehat{\boldsymbol{\Sigma}}_k^0\right)^{-1} \boldsymbol{S}_k$

$\boldsymbol{S}_k = \sum_{i=1}^k \boldsymbol{x}_i \widehat{e}_i, \quad k = 1, ..., n$

$\widehat{\boldsymbol{\Sigma}}_k^0$ — odhad varianční matice $\boldsymbol{S}_k$

$\widehat{e}_i,\ i = 1, \dots, n$ — rezidua vztažená k použitým odhadům parametru $\widehat{\beta}_k$

Výsledný test: H_0 zamítáme na hladině α:

$$T_n \ge t_n(\alpha) \qquad T_n(\varepsilon) \ge t_n(\alpha, \varepsilon)$$

$t_n(\alpha),\ t_n(\alpha, \varepsilon)$ — kritické hodnoty

Aproximace hodnot $t_n(\alpha)$ a $t_n(\alpha, \varepsilon)$ založeny na

(i) limitním rozdělení T_n popř. $T_n(\varepsilon)$ při H_0

(ii) resampling metody (bootstrap)

Odhady parametrů

• odhad bodu změny k^*:

$$\widehat{k}^* = argmax\Big\{(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0)^T \Big(\widehat{\boldsymbol{\Sigma}}_k\Big)^{-1}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_k^0);\ p < k < n-p\Big\}$$

analogicky definujeme $\widehat{k}_n^*(\varepsilon)$;

• $\widehat{\boldsymbol{\beta}}_{\widehat{k}^*}$ — odhad $\boldsymbol{\beta}$;

• $\widehat{\boldsymbol{\beta}}^0_{\widehat{k}^*}$ — odhad $\boldsymbol{\beta} + \boldsymbol{\delta}$.

Outline

Pravděpodobnostní nástroje

$X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s nulovou střední hodnotou a konečným rozptylem σ^2.

CLV

$S_n = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n X_i$ má "přibližně" normované normální rozdělení (ozn. $N(0,1)$), t.j. pro $n \to \infty$

$$S_n \to^d N, \quad N \quad má \quad N(0,1);$$

Pravděpodobnostní nástroje

$X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny s nulovou střední hodnotou a konečným rozptylem σ^2.

CLV

$S_n = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^{n} X_i$ má "přibližně" normované normální rozdělení (ozn. $N(0,1)$), t.j. pro $n \to \infty$

$$S_n \to^d N, \quad N \quad má \quad N(0,1);$$

FCLV, princip invariance

proces $\mathcal{S}_n = \{S_{\lfloor nt \rfloor};\ t \in [0,1]\}$ má "přibližně" rozdělení jako Brownův pohyb na $[0,1]$ (ozn. $\mathcal{W}$), tj. pro $n \to \infty$

$$Ef(\mathcal{S}_n) \to Ef(\mathcal{W})$$

pro každý omezený spojitý funkcionál;

Věta o extremálním rozdělení

(Darling- Erdős, 1956): je-li navíc $E|X_i|^{2+\delta} < \infty$ pro nějaké $\delta > 0$, $x \in R^1$

$$P\Big(\sqrt{2\log\log n}\max_{1\le k\le n}\frac{|S_k|}{\sqrt{k}\sigma} \le x + 2\log\log n - \frac{1}{2}\log\pi\Big) = \exp\{-e^{-x}\}.$$

Věta o extremálním rozdělení

(Darling- Erdős, 1956): je-li navíc $E|X_i|^{2+\delta} < \infty$ pro nějaké $\delta > 0$, $x \in R^1$

$$P\Big(\sqrt{2\log\log n}\max_{1\le k\le n}\frac{|S_k|}{\sqrt{k}\sigma} \le x + 2\log\log n - \frac{1}{2}\log\pi\Big) = \exp\{-e^{-x}\}.$$

Zákon iterovaného logaritmu

$$\limsup_{n\to\infty}\frac{1}{\sqrt{2n\log\log n}}|S_n| = \sigma, \quad a.s.$$

Věta o extremálním rozdělení

(Darling- Erdős, 1956): je-li navíc $E|X_i|^{2+\delta} < \infty$ pro nějaké $\delta > 0$, $x \in R^1$

$$P\Big(\sqrt{2\log\log n}\max_{1\le k\le n}\frac{|S_k|}{\sqrt{k}\sigma} \le x + 2\log\log n - \frac{1}{2}\log\pi\Big) = \exp\{-e^{-x}\}.$$

Zákon iterovaného logaritmu

$$\limsup_{n\to\infty}\frac{1}{\sqrt{2n\log\log n}}|S_n| = \sigma, \quad a.s.$$

KMT (Komlós, Major, Tusnády, 1975, 1976)

Můžeme definovat Brownův pohyb $\{W(t),\ t\in[0,\infty)\}$, že

$$S_n - W(n\sigma^2) = o(n^{1/(2+\delta)}), \quad a.s.$$

$$\max_{1\le k\le n}\frac{1}{k^{1/(2+\delta')}}|S_k - W(k\sigma^2)| = O_P(1) \text{ pro každé } \delta' > \delta\,.$$

Základní vlastnosti T_n a $T_n(\varepsilon)$

- regrese bez trendu , H_0 :

$$\lim_{n\to\infty} P\left(\sqrt{2\log\log n}\, T_n^{1/2} \le t + \frac{p}{2}\log\log\log n - \log(\Gamma(p/2))\right)$$

$$= \exp\{-2\exp\{-t\}\},\ t \in R^1;$$

$$T_n(\varepsilon) \to^d \sup_{\varepsilon<t<1-\varepsilon)} \left\{ \frac{\sum_{i=1}^p B_i^2(t)}{t(1-t)} \right\}$$

$\{B_j(t); t \in (0,1)\}$, $j = 1, ..., p$, — nezávislé Brownovy mosty $(B_i(t) = W_i(t) - tW_i(1))$

- regrese s trendem H_0:

$$T_n(\varepsilon) \to^d \sup_{\varepsilon<t<1-\varepsilon} \boldsymbol{S}^T(t,\boldsymbol{h})\Big(var\{\boldsymbol{S}(t,\boldsymbol{h})\}\Big)^{-1}\boldsymbol{S}(t,\boldsymbol{h})$$

$$\boldsymbol{S}(t,\boldsymbol{h}) = \int_0^t \boldsymbol{h}(x)dB_1(x) - \boldsymbol{C}(t)\boldsymbol{C}^{-1}(1)\int_0^1 \boldsymbol{h}(x)dB_1(x), \quad t \in [0,1];$$

Výsledky pro T_n zatím odvozeny jen pro polynomickou regresi, zobecnění předmětem výzkumu.

Jestliže změna nastala (H_1), testy založené na T_n nebo $T_n(\varepsilon)$ vedou ke správnému rozhodnutí s pravděpodobností blížící se k 1 (pro $n \to \infty$).

Jestliže $n\eta < k^* < (1-\eta)n$, pak $\widehat{k}_n^* - k^* = o_P(n)$.

Bootstrap

$$\begin{aligned} Y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + e_i, \quad i = 1\ldots, k^* \\ &= \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{x}_i^T \boldsymbol{\delta} + e_i, \quad i = k^* + 1\ldots, n, \end{aligned}$$

Chceme odhad rozdělení T_n při H_0.

Bootstrapový model:

$$Y_i^* = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n + e_i^*, \quad i = 1\ldots, n$$

$\widehat{\boldsymbol{\beta}}_n$ — odhad $\boldsymbol{\beta}$

$e_1^*, \ldots, e_n^*$– náhodný výběr z reziduí $\widehat{e}_1, \ldots, \widehat{e}_n$, popř. jejich permutací

Rozdělení odpovídající T_n^* popř. $T_n^*(\varepsilon)$ je pro široké spektrum situací dobrou aproximací pro rozdělení T_n popř. $T_n(\varepsilon)$ při H_0.

Outline

Ráztoka data-pokračování

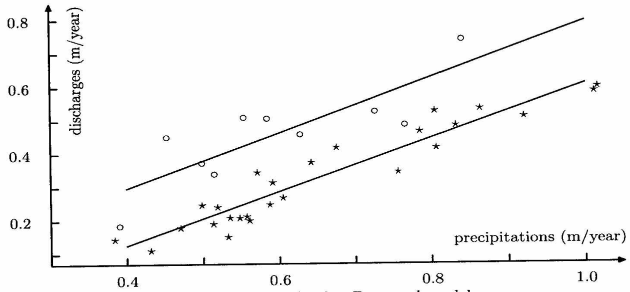
Modifikovaný test podílem věrohodností vedl k zamítnutí nulové hypotézy, tj. k závěru, že ke změně došlo

odhad bodu změny – po 26 letech

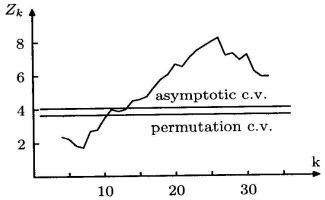
odhad parametrů před po změnou : -193.6   0.8

odhad parametrů po změně: -33.1   0.82

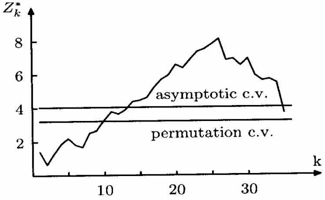
$T_n = \max_{1<k<n} Z_k$



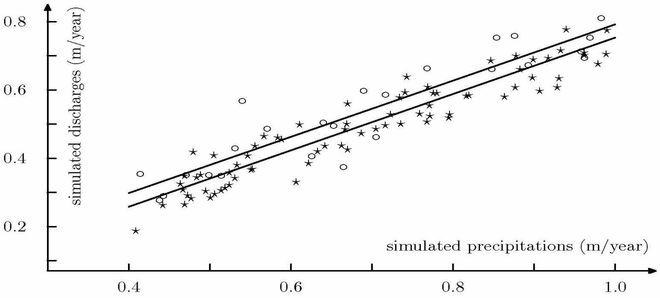
Malá Ráztoka: Data and model.



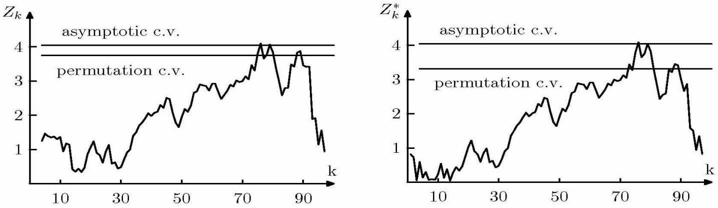
Statistics Z_k.



Statistics Z_k^*.



Simulated data and model.



Statistics Z_k.

Statistics Z_k^*.

Outline

Klementinská data

Průměrné roční teploty v Klementinu v letech 1775 – 1992

Zpracováváno mnoha statistiky:

Jarušková (1997)

Horváth a Kokoszka (2002)— neparametrická L_2 regrese

Gijbels a Goderniaux (2001)— neparametrická L_2 regrese

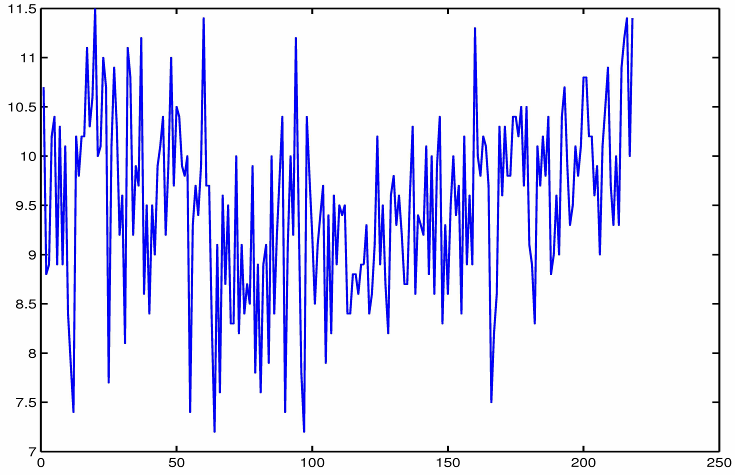
Janžura (2002) — metoda segmentace

Čapek a Hušková (2004)— neparametrická L_1 a L_2 regrese

Antoch (2006) — po částech lineární L_1 a L_2 regrese a MCMC

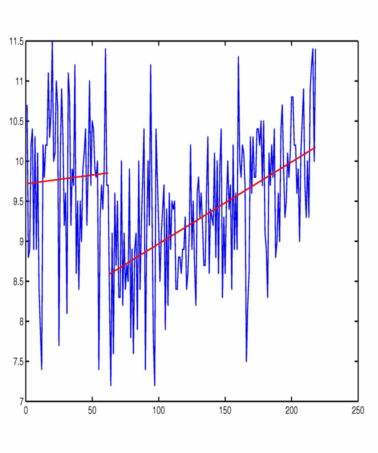
změny 1836, (1786, 1942)

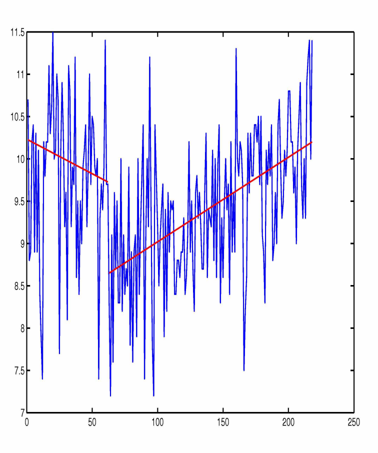
Data z Klementina



Data z Klementina

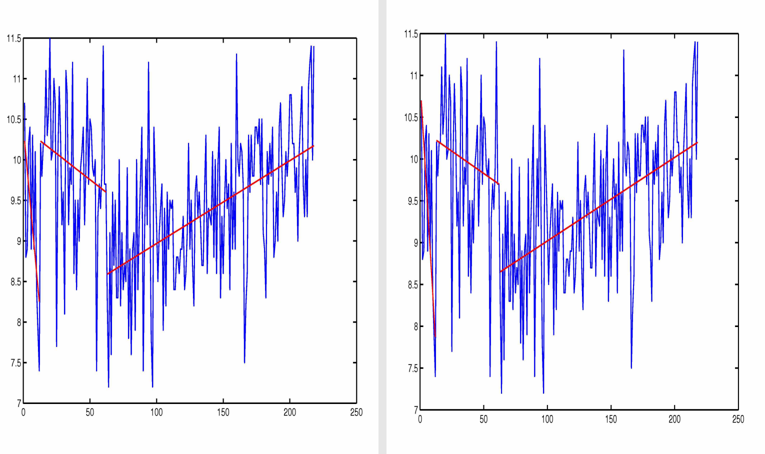
Po částech lineární regrese L_2 (vlevo) a L_1 (vpravo), Antoch, *1 změna*





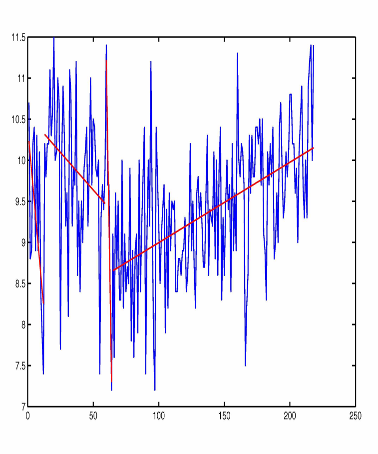
Data z Klementina

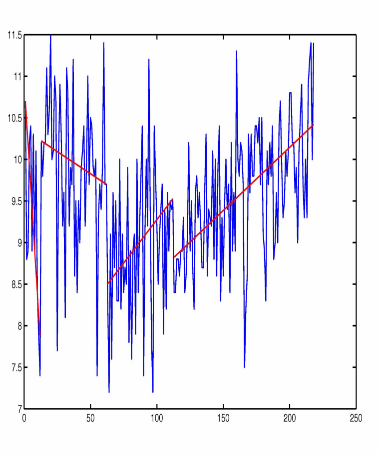
Po částech lineární regrese L_2 (vlevo) a L_1 (vpravo), Antoch, *2 změny*



Data z Klementina

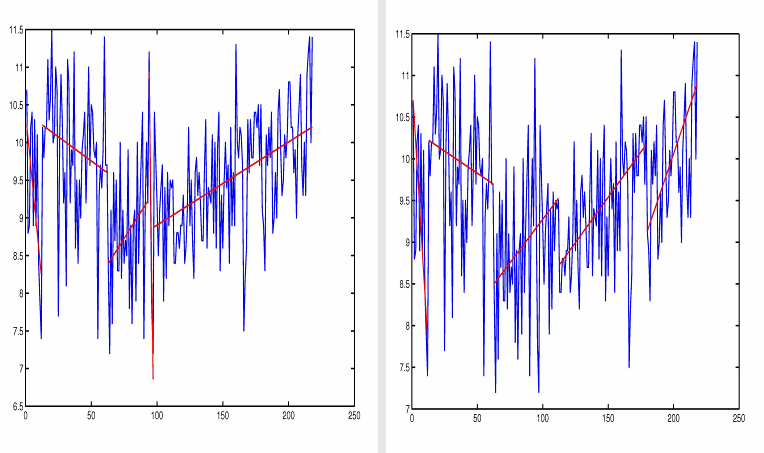
Po částech lineární regrese L_2 (vlevo) a L_1 (vpravo), Antoch, *3 změny*





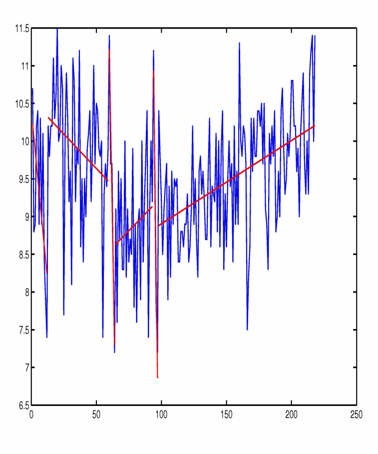
Data z Klementina

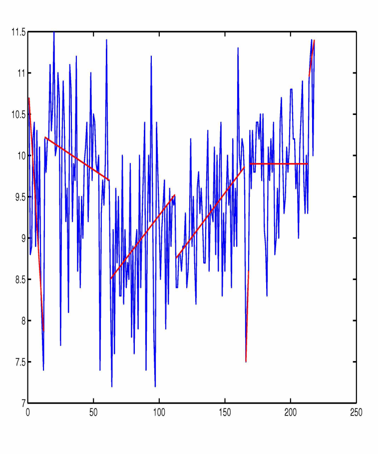
Po částech lineární regrese L_2 (vlevo) a L_1 (vpravo), Antoch, *4 změny*



Data z Klementina

Po částech lineární regrese L_2 (vlevo) a L_1 (vpravo), Antoch, *5 změn*





Neparametrická regrese — Klementinská data

Neparametrický model:

$$\begin{aligned} Y_i &= m_0(i/n) + e_i, \qquad & i = 1\dots, k^* \\ &= m_0(i/n) + \delta + e_i, \qquad & i = k^* + 1\dots, n, \end{aligned}$$

$m_0(.)$ — regresní hladká funkce na $[0,1]$

$m(x) = m_0(x) + \delta I\{x \geq k^*/n\}$ — regresní funkce

$\delta \neq 0$ — velikost skoku

$e_1, \dots, e_n$ — náhodné chyby (nezávislé stejně rozdělené náhodné veličiny)

2 odhady regresní funkce m v bodě x

$$\widehat{m}^{+}(x) = \sum_{i \geq nx} w_{in}(x) Y_i, \qquad \widehat{m}^{-}(x) = \sum_{i < nx} w_{in}(x) Y_i$$

$w_{in}(x)$ — váhy, např. jádrové odhady

$(\widehat{m}^{+}(x) = argmin\{\sum_{i \geq xn} (Y_i - a)^2 k((i/n - x)/h), a \in R^1\})$

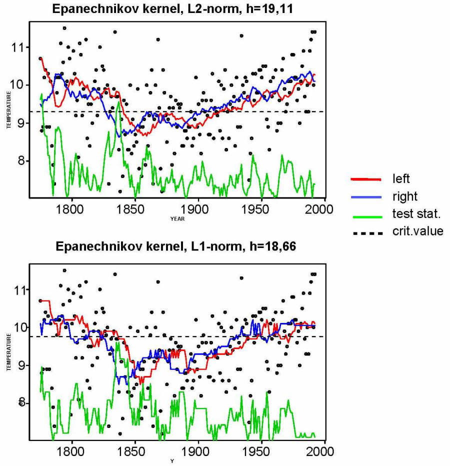
J

e-li m hladká v bodě x, pak $\widehat{m}^{+}(x) - \widehat{m}^{-}(x) \approx 0$.

Má-li m skok v x, pak $|\widehat{m}^{+}(x) - \widehat{m}^{-}(x)| \approx |\delta|$.

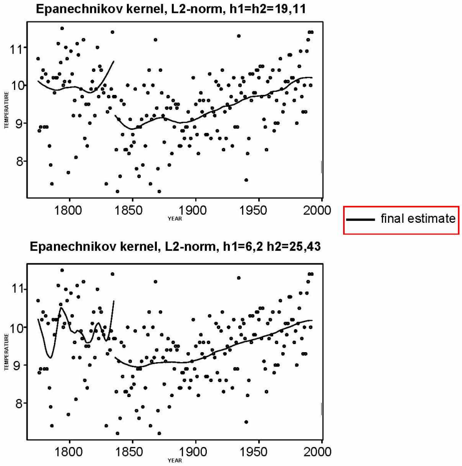
V následující grafech jsou $\widehat{m}^{+}(x)$ a $\widehat{m}^{-}(x)$ a vhodně standardizovaná $|\widehat{m}^{+}(x) - \widehat{m}^{-}(x)|$

Neparametrický přístup



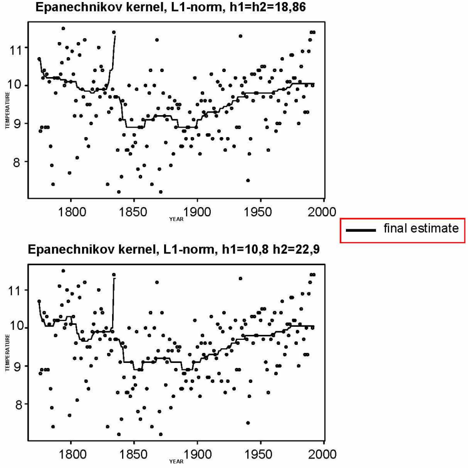
Klementinum data - Epanechnikov kernel

Neparametrický přístup



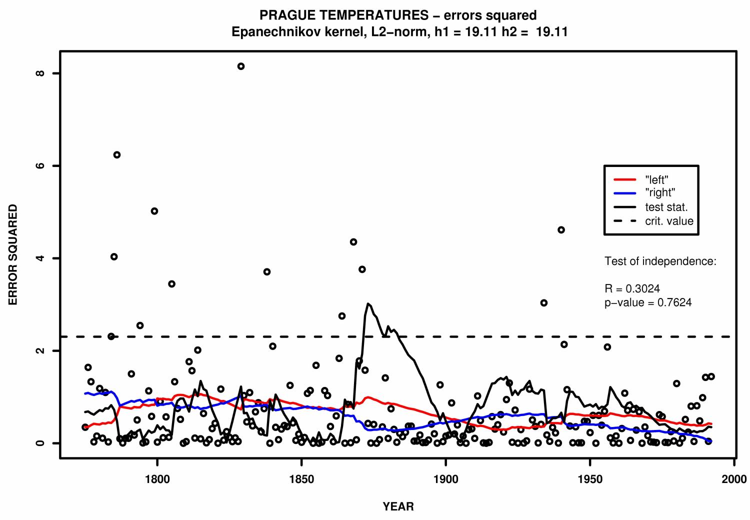
Klementinum data - final estimates, Epanechnikov kernel, L2-norm

Neparametrický přístup



Klementinum data - final estimates,
Epanechnikov kernel, L1-norm

Neparametrický přístup



Poznámky

• K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.

Poznámky

• K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.
• VÝPOČETNI ASPEKTY!!!! Postupy jsou výpočetně náročné.

Poznámky

• K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.
• VÝPOČETNI ASPEKTY!!!! Postupy jsou výpočetně náročné.

• Předpoklady na náhodné chyby $e_1, \ldots, n$, nezávislé stejně rozdělené není v často splněno. Problém co nejslabších předpokladů.

Poznámky

• K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.
• VÝPOČETNI ASPEKTY!!!! Postupy jsou výpočetně náročné.

• Předpoklady na náhodné chyby $e_1, \dots, n$, nezávislé stejně rozdělené není v často splněno. Problém co nejslabších předpokladů.

• L_2, L_1 postupy. L_2 — citlivé na odlehlá pozorování, výjimečná pozorování tak pak indikují změnu. L_1 vhodnější.

Poznámky

- K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.
- VÝPOČETNI ASPEKTY!!!! Postupy jsou výpočetně náročné.
- Předpoklady na náhodné chyby $e_1, \ldots, n$, nezávislé stejně rozdělené není v často splněno. Problém co nejslabších předpokladů.
- L_2, L_1 postupy. L_2 — citlivé na odlehlá pozorování, výjimečná pozorování tak pak indikují změnu. L_1 vhodnější.
- Časové řady — v řadě případů podobné regresi, v současné době pozornost změnám typu závislosti, hlavně časové finanční řady.

Poznámky

- K datům je potřeba navrhnout model co nejjednodušší, ale výstižný. Je třeba studovat teoretické vlastnosti, provést simulační studii, aplikovat na data.
- VÝPOČETNI ASPEKTY!!!! Postupy jsou výpočetně náročné.
- Předpoklady na náhodné chyby $e_1, \dots, n$, nezávislé stejně rozdělené není v často splněno. Problém co nejslabších předpokladů.
- L_2, L_1 postupy. L_2 — citlivé na odlehlá pozorování, výjimečná pozorování tak pak indikují změnu. L_1 vhodnější.
- Časové řady — v řadě případů podobné regresi, v současné době pozornost změnám typu závislosti, hlavně časové finanční řady.
- Vícenásobné změny — speciální postupy.

- Postupné změny (speciální případ regrese s trendem – Jarušková, problematika klimatologie a příbuzných oblastí).
- Postupy i pro další situace.
- Sekvenční verze, tzv. monitoring při trénovacích datech.

Změna v průměru versus náhodná procházka a změna v průměru versus long range dependence

Model polohy se změnou

$$Y_i = \mu_i + e_i, \quad i = 1 \ldots, n,$$

$\mu_1, \ldots, \mu_n$ — nenáhodná část (poloha, střední hodnota, medián)

$e_1, \ldots, e_n$ — náhodná část (náhodné chyby)

$H_0 : \mu_1 = \ldots = \mu_n$ versus
$H_1 : \mu_1 = \ldots = \mu_{k^*} \neq \mu_{k^*+1} = \ldots = \mu_n$ for some $k^* < n$

$k^* = \lfloor n\theta \rfloor$ pro nějaké $\theta \in (0, 1)$

rozdělení náhodných chyb:

(I) slabá závislost: existuje $\sigma > 0$, že pro $n \to \infty$

$$\Big\{\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nt\rfloor} e_i,\, t\in[0,1]\Big\} \to^{\mathcal{D}[0,1]} \sigma\Big\{W(t),\ t\in[0,1]\Big\}$$

$\{W(t); t\in[0,1]\}$ je Brownův pohyb;

rozdělení náhodných chyb:

(I) slabá závislost: existuje $\sigma > 0$, že pro $n \to \infty$

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} e_i,\, t \in [0,1] \right\} \to^{\mathcal{D}[0,1]} \sigma \left\{ W(t),\ t \in [0,1] \right\}$$

$\{W(t); t \in [0,1]\}$ je Brownův pohyb;

(II) náhodná procházka: existuje $\sigma > 0$, že pro $n \to \infty$,

$$\left\{ \frac{1}{\sqrt{n}} e_{\lfloor nt \rfloor},\, t \in [0,1] \right\} \to^{\mathcal{D}[0,1]} \sigma \left\{ W(t), t \in [0,1] \right\};$$

rozdělení náhodných chyb:

(I) slabá závislost: existuje $\sigma > 0$, že pro $n \to \infty$

$$\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} e_i, \, t \in [0,1] \right\} \to^{\mathcal{D}[0,1]} \sigma \left\{ W(t), \; t \in [0,1] \right\}$$

$\{W(t); t \in [0,1]\}$ je Brownův pohyb;

(II) náhodná procházka: existuje $\sigma > 0$, že pro $n \to \infty$,

$$\left\{ \frac{1}{\sqrt{n}} e_{\lfloor nt \rfloor}, \, t \in [0,1] \right\} \to^{\mathcal{D}[0,1]} \sigma \left\{ W(t), t \in [0,1] \right\};$$

(III) long range dependence: existuje $\sigma > 0$, že pro $n \to \infty$,

$$\left\{ \frac{1}{n^{2H} L(n)} \sum_{i=1}^{\lfloor nt \rfloor} e_i, \, t \in [0,1] \right\} \to^{\mathcal{D}[0,1]} \sigma \left\{ W_H(t), \; t \in [0,1] \right\}$$

$\{W_H(t); t \in [0,1]\}$ je frakcionální Brownův pohyb s $1/2 < H < 1$ ($E\, W_H(t_1) W_H(t_2) = (t_1^{2H} + |t_2^{2H} - |t_1 - t_2|^{2H})/2$, $0 < t_1, t_2 < 2$)

Statistické modely změn

Outline

Úvod

Regrese

Pravděpodob nástroje

Raztoka

Klementinum

Finanční časové řady

Proces částečných součtů $S_n(t) = \sum_{i=1}^{\lfloor nt \rfloor} e_i, \quad t \in (0,1)$

$$\text{(I): } \{\frac{1}{\sigma\sqrt{n}} S_n(t);\ t \in (0,1)\} \to^{\mathcal{D}[0,1]} \{W(t);\ t \in (0,1)\}$$

$$\text{(II): } \{\frac{1}{\sigma n^{3/2}} S_n(t);\ t \in (0,1)\} \to^{\mathcal{D}[0,1]} \{\int_0^t W(s)ds;\ t \in (0,1)\}$$

$$\text{(III): } \left\{\frac{1}{\sigma n^H \sqrt{L(n)}} \sum_{i=1}^{\lfloor nt \rfloor} e_i, t \in (0,1)\right\} \to^{\mathcal{D}[0,1]} \left\{W_H(t),\ t \in (0,1)\right\}$$

$1/2/H < 1$

Testová statistika

$$T^0_{n,1} = \max_{1\le k\le n} \frac{1}{\sqrt{n}} \Big| \sum_{i=1}^{k} (Y_i - \overline{Y}_n) \Big|,$$

$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$

- H_0 a (I), pak $T^0_{n,1} = O_P(1)$
- H_0 a(II) nebo (III), $T^0_{n,1} \to^P \infty$
- H_1 a (I), pak $T^0_{n,1} \to^P \infty$

Tudíž můžeme zamítnout hypotézu H_0, i když platí.

Navržen dvoustupňový postup:

(A) aplikujeme obvyklý test pro H_0 versus H_1 při (I), nevede-li k zamítnutí H_0, ukončíme postup;

Navržen dvoustupňový postup:

(A) aplikujeme obvyklý test pro H_0 versus H_1 při (I), nevede-li k zamítnutí H_0, ukončíme postup;

(B) Vede-li k zamítnutí, data rozdělíme na dvě skupiny $Y_1, \dots, Y_{\widehat{k}^*}$ respektive $Y_{\widehat{k}^*+1}, \dots, Y_n$, kde $\widehat{k}^*$ je odhad bodu změny k^*. Test pak aplikujeme na na jednotlivé skupiny. Pokud ani jeden test nevede k zamítnutí, rozhodujeme se pro H_1 s odhadem bodu změny $\widehat{k}^*$.

Podrobně studován případ s testovou statistikou $T_{n,1}/s_n$

s_n^2 — odhad σ^2 (modifikovaný Bartlettův odhad) při (I), tj. při (I) $s_n^2 \to^P \sigma^2$

$\widehat{k}^* = argmax\Big\{k = 1, \dots, n; \Big|\sum_{i=1}^{k}(Y_i - \overline{Y}_n)\Big|\Big\}$

Při H_0, (I) a je-li s_n^2 konzistentní odhad σ^2 platí

$$\lim_{n\to\infty} P(T_{n,1}/s_n \le x) = P(sup_{0<t<1}|B(t)| \le x), \quad x \in R^1$$

$\{B(t),\ t \in (0,1)\}$ — Brownův most

Aplikace na časové finanční řady

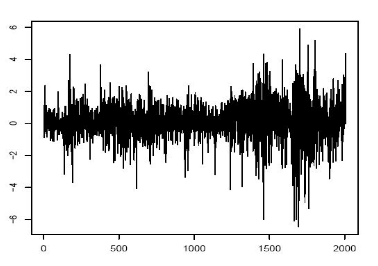
Denní výnos indexu DAX 30 (německý akciový trh 2.1.1992 – 31.12.1999

P_i — výnos v den i, $i = 1, \dots, n = 2009$

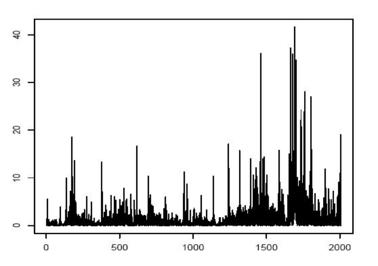
$Q_i = 100 \log(P_i / P_{i-1}),\ i = 1, \dots, n$
na základě $T^0_{1,n}(\boldsymbol{Q})$ nelze zamítnout hypotézu o konstantnosti středních hodnot

$Y_i = Q_i^2,\ i = 1, \dots, n$
na základě $T^0_{1,n}$ zamítáme hypotézu o konstantnosti volatility, max dosaženo pro $\widehat{k}^* = 1385$

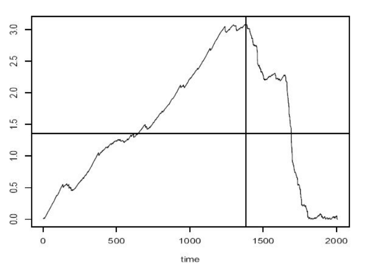
data rozdělíme na dvě skupiny, prvních 1385 a ostatní, na každou skupinu aplikujeme $T^0_{1,n}$
závěr: změna ve volatilitě nastala, $\widehat{k}^* = 1385$



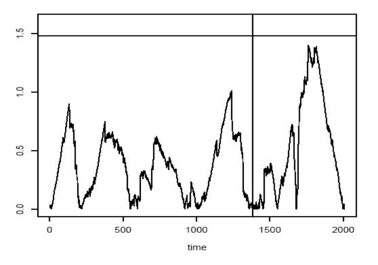
log returns



Squared log returns



$T^0_{1,n}$

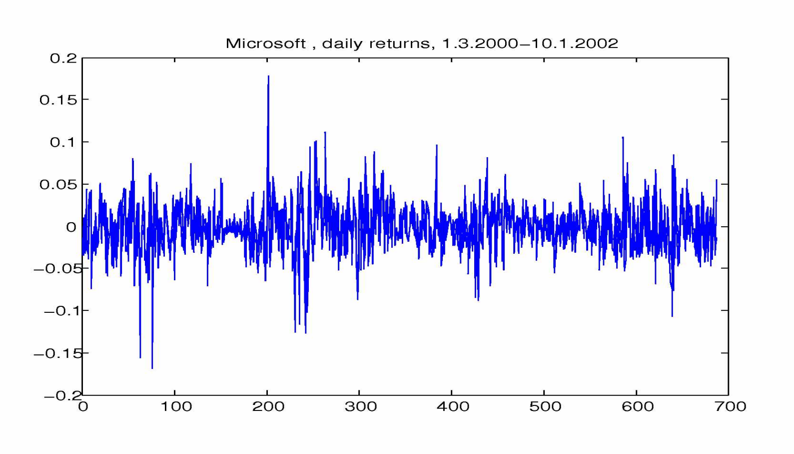


modif. $T^0_{1,n}$

Outline

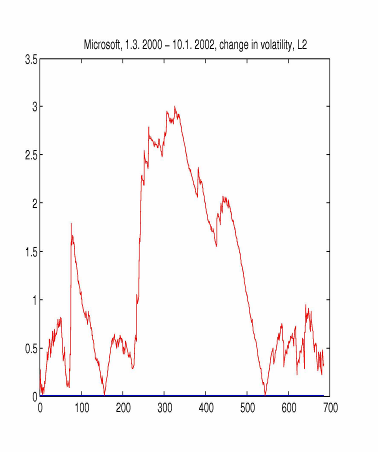
Další aplikace na finanční časové řady

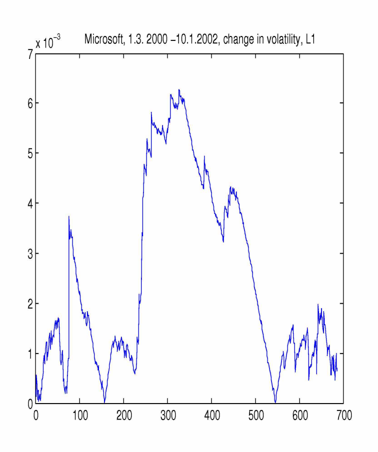
(i) *Denní výnos ceny akcií obchodovaných ve dnech 1.3.2000-10.1. 2002 (data ze zahraničních pramenů).*



Data lze Box-Jenkinsonovou metodologií identifikovat jako bílý v šum

Test změny ve volatilitě (regrese na druhé mocniny výnosu) v L_2 a L_1 změnu neprokázal (T_n, kritická hodnota 3, 6953).

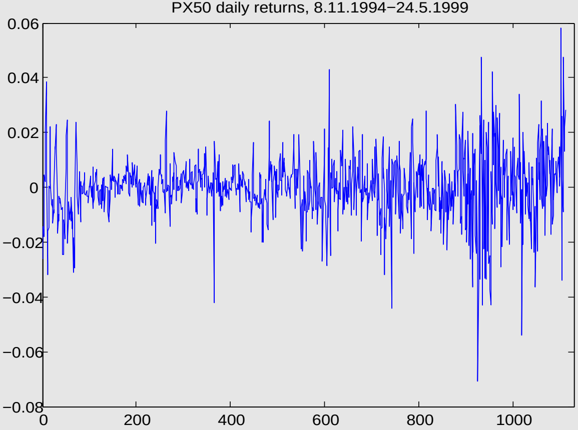




Další analýza vedla k závěru: stacionární proces $GARCH(1,1)$.

(ii) *Denní výnos indexu PX50 ve dnech 8.11. 1994 - 24.5. 1999*

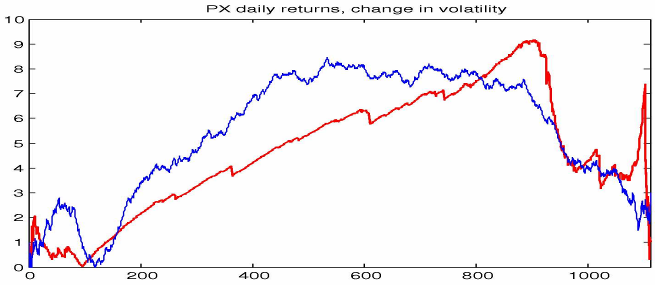
Data lze Box-Jenkinsonovou metodologií identifikovat jako AR(1) proces s náhodným koeficientem, jehož chybový proces není stacionární



Test změny ve volatilitě v modelu

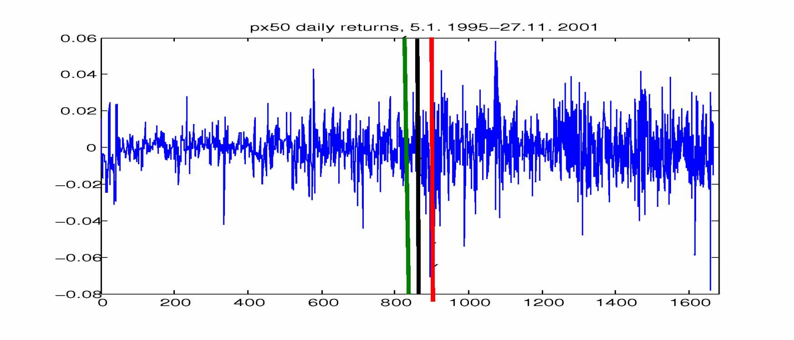
$$Y_i = \beta Y_{i-1} + e_i, \qquad Ee_i^2 | Y_{i-1}^2 = \alpha_0 + \alpha_1 Y_{i-1}^2$$

červená — výsledky pro L_2, modrá — výsledky pro L_1
L_2: 9.164 — hodnota testové statistiky $\sqrt{T_n}$, $k^* = 901$ (29.5. 1998)
L_1: 8.4639 — hodnota testové statistiky $\sqrt{T_n}$, $k^* = 533$



Oba postupy prokazují změnu (kritická hodnota 3, 7) .

(iii) *Denní výnos indexu PX50 ve dnech 5.1.1995 - 27.11. 2001.*

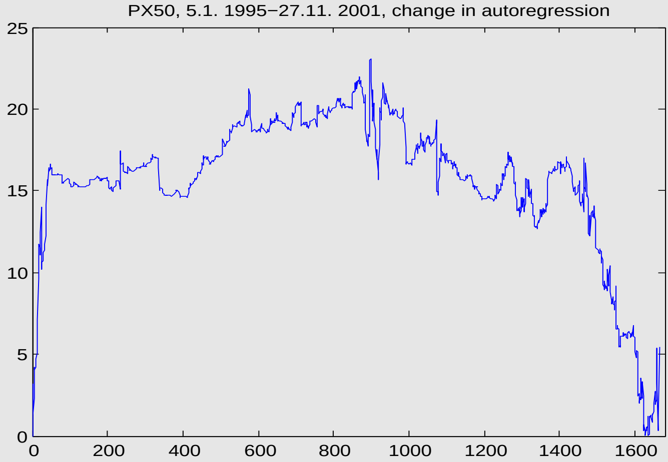


Test změny v autoregresi 1. řádu, změna indikována

—- bod změny v autoregresi, (1.9. 1998)
—- bod změny ve volatilitě z dat 8.11.1994 - 24.5.1999
—- bod změny ve volatilitě 5.1.1995 - 27.11. 2001 (18.6. 1998)

Statistické modely změn

Outline
Úvod
Regrese
Pravděpodob nástroje
Raztoka
Klementinum
Finanční časové řady


PX50, 5.1. 1995−27.11. 2001, change in autoregression
25
20
15
10
5
0
0
200
400
600
800
1000
1200
1400
1600