Testování změn v binárnách autoregresních modelech

Šárka Hudecová

KPMS MFF UK

ROBUST 2012

Němčičky 9.–14.9.2012

Uvažovaná situace

Časová řada binárních veličin

- posloupnost $\{Y_t\}$, kde Y_t nabývá hodnot 0 nebo 1
  - ↪ výskyt nějaké události v čase
  - ↪ denní výskyt srážek, ukazatel recese aj.

Uvažovaná situace

Časová řada binárních veličin

- posloupnost $\{Y_t\}$, kde Y_t nabývá hodnot 0 nebo 1
  - $\hookrightarrow$ výskyt nějaké události v čase
  - $\hookrightarrow$ denní výskyt srážek, ukazatel recese aj.
- různé modely a přístupy k analýze
  - $\hookrightarrow$ snaha o analogii ARMA modelů $\leftrightsquigarrow$ více možností, komplikovanější
  - $\hookrightarrow$ binární autoregresní modely BAR $\rightsquigarrow$ časové řady řídící se zobecněným lineárním modelem

Uvažovaná situace

Časová řada binárních veličin

- posloupnost $\{Y_t\}$, kde Y_t nabývá hodnot 0 nebo 1
  - ↪ výskyt nějaké události v čase
  - ↪ denní výskyt srážek, ukazatel recese aj.
- různé modely a přístupy k analýze
  - ↪ snaha o analogii ARMA modelů $\leftrightsquigarrow$ více možností, komplikovanější
  - ↪ binární autoregresní modely BAR $\rightsquigarrow$ časové řady řídící se zobecněným lineárním modelem
- možná změna v modelu v neznámém okamžiku
  - ↪ změna v parametrech
  - ↪ test, detekce změny

Obsah

Obsah

Výskyt srážek

- denní výskyt srážek v oblasti Mokré louky v Jižních Čechách z období 1977-2011 (12 783 záznamů)
- domněnka, že v posledních letech nastala změna ve výskytu

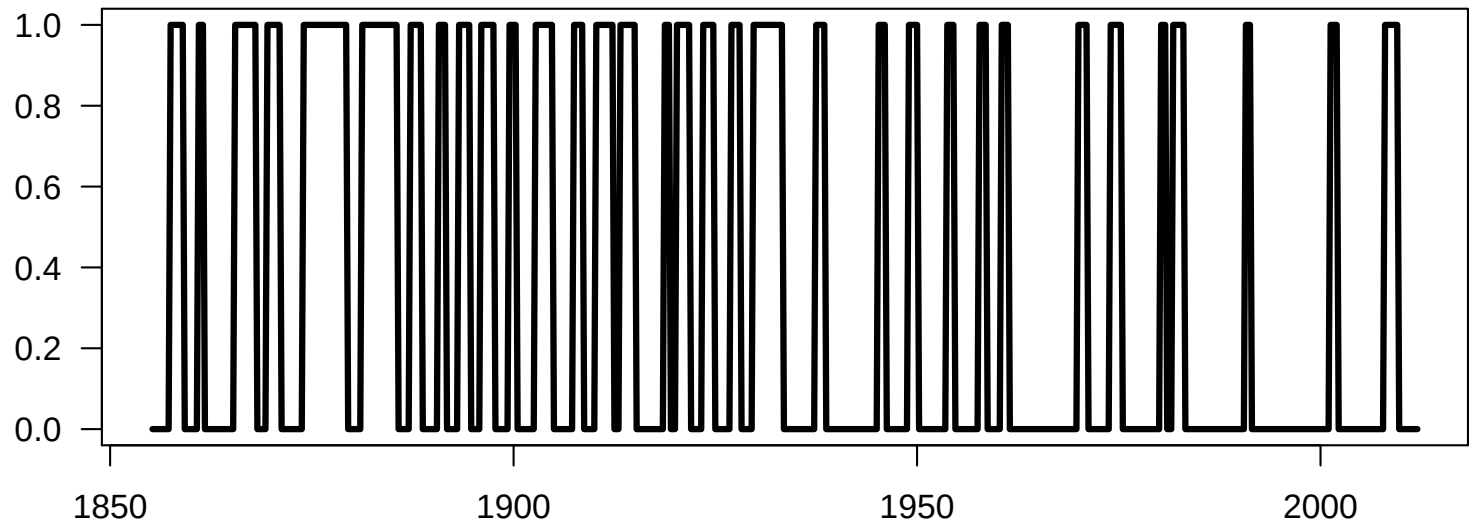
Výskyt srážek

- denní výskyt srážek v oblasti Mokré louky v Jižních Čechách z období 1977-2011 (12 783 záznamů)
- domněnka, že v posledních letech nastala změna ve výskytu

- standardní (historický) přístup $\rightsquigarrow$ Markovské řetězce, speciálně MŘ 1. řádu
- chceme-li brát v úvahu vysvětlující proměnné (měsíc, teplota apod.) $\rightsquigarrow$ BAR modely
- zajímá nás, zda došlo k nějaké změně

Index recese USA

- čtvrtletní indikátor recese v USA z let 1855–2011 (628 záznamů)
- změna kolem roku 1945, tj. kolem 2. světové války



Obsah

Binární autoregresní modely

↪ $\{Y_t\}$ binární časová řada, vysvětlující proměnné $\{X_t, W_t\}$

↪ informace známá do času $t-1$

$$\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots, W_t, W_{t-1}, \ldots\}$$

↪ model pro $\pi_t = \mathrm{P}(Y_t = 1|\mathcal{F}_{t-1}) = \mathrm{E}[Y_t|\mathcal{F}_{t-1}]$:

$$\mathrm{logit}[\pi_t] = \beta' \mathbf{Z}_{t-1},$$

kde

- $\mathbf{Z}_{t-1}$ je q-dimenzionální vektor obsahující minulé hodnoty $\{Y_t\}$ a vysvětlujících proměnných $\{X_t, W_t\}$
- β je vektor neznámých parametrů

Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis.* Wiley, New York.

Binární autoregresní modely

Příklad: BAR(p) bez vysvětlujících proměnných

$$\hookrightarrow \quad \mathcal{F}_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \dots\}$$

$$\hookrightarrow \quad Z_{t-1} = (1, Y_{t-1}, \dots, Y_{t-p})'$$

$$\hookrightarrow \quad \text{logit}\,[\pi_t] = \beta_0 + \beta_1 Y_{t-1} + \dots \beta_p Y_{t-p}$$

Poznámka

- existují i jiné přístupy (modifikace) k modelování binárních (neGaussovských) časových řad

Odhad parametrů

Odhad parametrů $\leftrightsquigarrow$ metodou partial maximum likelihood

- partial likelihood

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{t=1}^{n} [\pi_t(\boldsymbol{\beta})]^{Y_t} [1 - \pi_t(\boldsymbol{\beta})]^{1-Y_t},$$

- odhad (MPLE) $\widehat{\boldsymbol{\beta}}$ je bod maxima $\mathrm{PL}(\boldsymbol{\beta})$, tj. řešení $\nabla \log \mathrm{PL}(\boldsymbol{\beta}) = 0$,
- numerické řešení

$$\sum_{t-1}^{n} \mathbf{Z}_{t-1}[Y_t - \pi_t(\boldsymbol{\beta})] = 0,$$

- odhad v praxi ➡ R funkce `glm`

Vlastnosti odhad parametrů

Za určitých předpokladů regularity platí

1. $\widehat{\beta}$ je s.j. jediný pro dostatečně velké n,
2. $\widehat{\beta}$ je konzistentní odhad β a

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathrm{N}(0, G^{-1}(\beta)), \quad n \to \infty,$$

kde G je limitní informační matice (nenáhodná)

$$\frac{G_n(\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n} \sigma_t^2 \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} \to G(\beta), \quad n \to \infty, \quad \text{s.j.}$$

Vlastnosti odhad parametrů

Za určitých předpokladů regularity platí

1. $\widehat{\beta}$ je s.j. jediný pro dostatečně velké n,
2. $\widehat{\beta}$ je konzistentní odhad β a

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathrm{N}(0, G^{-1}(\beta)), \quad n \to \infty,$$

kde G je limitní informační matice (nenáhodná)

$$\frac{G_n(\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n} \sigma_t^2 \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} \to G(\beta), \quad n \to \infty, \quad \text{s.j.}$$

Testování hypotéz

- založeno na partial likelihood
- podíl věrohodností, Waldův test, skórový test

Obsah

Model se změnou

- data: n realizací $(Y_t, Y_{t-1}, \ldots, Y_{t-p})'$
- model se změnou

$$\text{logit}(\pi_t) = \begin{cases} \beta_0 + \sum_{j=1}^{p} \beta_j Y_{t-j} = \boldsymbol{\beta}' \mathbf{Z}_{t-1}, & t = 1, \ldots, m \\ \beta_0^* + \sum_{j=1}^{p} \beta_j^* Y_{t-j} = \boldsymbol{\beta}^{*\prime} \mathbf{Z}_{t-1}, & t = m+1, \ldots, n, \end{cases}$$

  kde $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$
- test, zda nastala změna, tj.

$$H_0 : m = n \quad \text{proti } H_1 : m < n$$

Vliv změny v parametrech — BAR(1) model

Příklad: BAR(1)

$$\text{logit}\,[\pi_t] = \beta_0 + \beta_1 Y_{t-1}$$

- $\{Y_t\}$ je Markovský řetězec 1.řádu s pstmi přechodů

$$p_{i1} = \frac{1}{1 + \mathrm{e}^{(-\beta_0 - \beta_1 i)}}, \quad p_{i0} = 1 - p_{i1}, \quad i = 0, 1,$$

- změna v β_0 implikuje změnu v p_{ij} pro všechna $i, j = 0, 1$; změna v β_1 má vliv jen na p_{11} a p_{10}
- změna v β_0 nebo $\beta_1 \rightsquigarrow$ změna ve stacionárním rozdělení

Detekce změn v GLM

- change-point v klasických lineárních modelech, ARMA modelech

  
  Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* Wiley, New York.

  
  články prof. Huškové, doc. Práškové, prof. Antocha, prof. Jarůškové a další

- change-point v GLM

  
  Antoch, J., Gregoire, G., and Jarůšková, D. (2004). Detection of structural changes in generalized linear models. *Stat. and Probab. Lett.*, 69:315–332.

  ↪ testová statistika ⇜⇝ maximum skórových statistik
  ↪ za H_0 asymptoticky Darling-Erdös typ limitního rozdělení

Model se změnou v absolutním členu

Budeme uvažovat model

$$\text{logit}(\pi_t) = \begin{cases} \beta_0 + \sum_{j=1}^{p} \beta_j Y_{t-j}, & t \leq m \\ \beta_0^* + \sum_{j=1}^{p} \beta_j Y_{t-j}, & t = m+1, \ldots, n, \end{cases}$$

kde $\beta_0 \neq \beta_0^*$, a test $H_0 : m = n$ proti $H_1 : m < n$

- $\hookrightarrow$ odvodíme test pro tuto situaci
- $\hookrightarrow$ simulace $\rightsquigarrow$ lze úspěšně použít i v obecnější situaci (změna nejen v interceptu)

Model se změnou v absolutním členu

Budeme uvažovat model

$$\text{logit}(\pi_t) = \begin{cases} \beta_0 + \sum_{j=1}^{p} \beta_j Y_{t-j}, & t \leq m \\ \beta_0^* + \sum_{j=1}^{p} \beta_j Y_{t-j}, & t = m+1, \ldots, n, \end{cases}$$

kde $\beta_0 \neq \beta_0^*$, a test $H_0 : m = n$ proti $H_1 : m < n$

Odvození testové statistiky

- $m = k$ známé $\rightsquigarrow$ test založený na skórové statistice $c_n^{(k)}$
- m neznámé $\rightsquigarrow$ za testovou statistiku T_n vezmeme maximum $\sqrt{c_n^{(k)}}$ přes všechna možná k
- ukážeme, že asymptotické rozdělení T_n stejné jako v GLM — článek Antoch et al. (2004)

Testová statistika

Značení

$\hookrightarrow$ $\widehat{\pi}_t$ je odhad $\pi_t = \mathrm{P}(Y_t = 1 | \mathcal{F}_{t-1})$ za H_0,

$\hookrightarrow$ $\widehat{S}_k = \sum_{t=1}^{k} (Y_t - \widehat{\pi}_t)$,

$\hookrightarrow$

$$V_k = \sum_{t=1}^{k} \widehat{\sigma}_t^2 - \left[\sum_{t=1} \widehat{\sigma}_t^2 \mathbf{Z}_{t-1}\right]' \left(\sum_{t=1}^{n} \widehat{\sigma}_t^2 \mathbf{Z}_{t-1} \mathbf{Z}_{t-1}'\right)^{-1} \left[\sum_{t=1} \widehat{\sigma}_t^2 \mathbf{Z}_{t-1}\right],$$

kde $\mathbf{Z}_{t-1} = (1, Y_{t-1}, \ldots, Y_{t-p})'$ a $\widehat{\sigma}_t^2 = \widehat{\pi}_t(1 - \widehat{\pi}_t)$

Testová statistika

$$T_n = \max_{k_0 \le k \le n - k_0} \frac{|\widehat{S}_k|}{\sqrt{V_k}},$$

kde k_0 je takové, že $V_k > 0$ pro všechna $k_0 < k < n - k_0$.

Předpoklady

(A1) Skutečná hodnota β leží v otevřené podmnožině $\mathbb{R}^{p+1}$.

(A2) Vektor $\mathbf{Z}_{t-1}$

$\hookrightarrow$ leží s.j. v nenáhodné kompaktní $\Lambda \subset \mathbb{R}^{p+1}$ a

$\hookrightarrow$ $\sum_{t=1}^{n} \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} > 0$ s pravděpodobností 1.

(A3) Existuje limitní informační matice $G(\beta) > 0$ taková, že

$$\frac{G_n(\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n} \sigma_t^2 \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} \to G(\beta), \quad n \to \infty, \quad \text{s.j.}$$

(A4) S pravděpodobností 1 platí

$$\left\| \frac{1}{k}\sum_{t=1}^{k} \sigma_t^2 \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} - G \right\| = o\left(\frac{1}{\log k}\right) \text{ pro } k \to \infty,$$

$$\left\| \frac{1}{n-k}\sum_{t=k+1}^{n} \sigma_t^2 \mathbf{Z}_{t-1}\mathbf{Z}'_{t-1} - G \right\| = o\left(\frac{1}{\log(n-k)}\right), \text{ pro } k, n-k \to \infty.$$

Rozdělení testové statistiky

Věta

Za platnosti (A1)–(A4) má za nulové hypotézy $H_0 : m = n$

testová statistika $T_n = \max_{k_0 \le k \le n-k_0} \frac{|\widehat{S}_k|}{\sqrt{V_k}}$ *asymptotické rozdělení*

$$\mathrm{P}\left(T_n < \sqrt{2 \log\log n} + \frac{\log\log\log n}{2\sqrt{2\log\log n}} + \frac{t - \frac{1}{2}\log\pi}{\sqrt{2\log\log n}} \right) \to e^{-2e^{-t}}$$

pro $t \in \mathbb{R}$ *a* $n \to \infty$.

Postup důkazu

- $S_k = \sum_{t-1}^{k}(Y_t - \pi_t) = \sum_{t=1}^{k} X_t$ je martingal:
  - ↪ martingalové diference $\{X_t\}$ omezené,
  - ↪ $s_n^2 = \sum_{t=1}^{n} \sigma_t^2$ splňuje $s_n^2 \to \infty$ s.j. pro $n \to \infty$, spec.

$$\frac{s_n^2}{n} - A = o\left(\frac{1}{\log n}\right) \quad \text{s.j.},$$

    kde $A > 0$ je (1,1) prvek matice G

  - ↪ vlastnosti $\{S_k\}$
- Antoch et al. (2004): rozdělení $T_n = \max_{1 \le k \le n} \frac{|\widehat{S}_k|}{\sqrt{V_k}}$ ⇜⇝ asymptoticky stejné jako rozdělení

$$\max_{1 \le k \le n} \sqrt{\frac{n}{k(n-k)A}} \left| S_k - \frac{k}{n} S_n \right|$$

  - ↪ využije se vnoření $\{S_n\}$ do Wienerova procesu
  - ↪ odvození rozdělení pomocí Brownova mostu

Poznámky

- jestliže $T_n = \max_{1 \le k \le n} \frac{|\widehat{S}_k|}{\sqrt{V_k}} > c_\alpha \leftrightsquigarrow$ argmax odhad m
- $T_n \leftrightsquigarrow$ test proti alternativě změny v absolutním členu
  - $\hookrightarrow$ simulace $\rightsquigarrow$ funguje dobře o proti složitější alternativě
- další možné testové statistiky

$$U_n = \max_{k_0 < k < n-k_0} \left\{ \sqrt{\frac{\sum_{t=1}^{n} \widehat{\sigma}_t^2}{\sum_{t=1}^{k} \widehat{\sigma}_t^2 \sum_{t=k+1}^{n} \widehat{\sigma}_t^2}} \cdot |\widehat{S}_k| \right\}$$

$$W_n = \max_{k_0 < k < n-k_0} \left\{ \sqrt{\frac{n}{k(n-k)}} \frac{|\widehat{S}_k|}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} \widehat{\sigma}_t^2}} \right\}$$

  - $\hookrightarrow$ výpočetně jednodušší
  - $\hookrightarrow$ simulace $\rightsquigarrow$ menší síla

Obsah

Simulace

Chování T_n

- za H_0
  - ↪ přesnost aproximace pro konečné n
  - ↪ dodržování hladiny testu
- za H_1
  - ↪ závislost síly testu na různých parametrech změny
  - ↪ síla proti obecnější alternativě (změna nejen v abs. členu)

Simulace

Chování T_n

- za H_0
  - ↪ přesnost aproximace pro konečné n
  - ↪ dodržování hladiny testu
- za H_1
  - ↪ závislost síly testu na různých parametrech změny
  - ↪ síla proti obecnější alternativě (změna nejen v abs. členu)

Výsledky

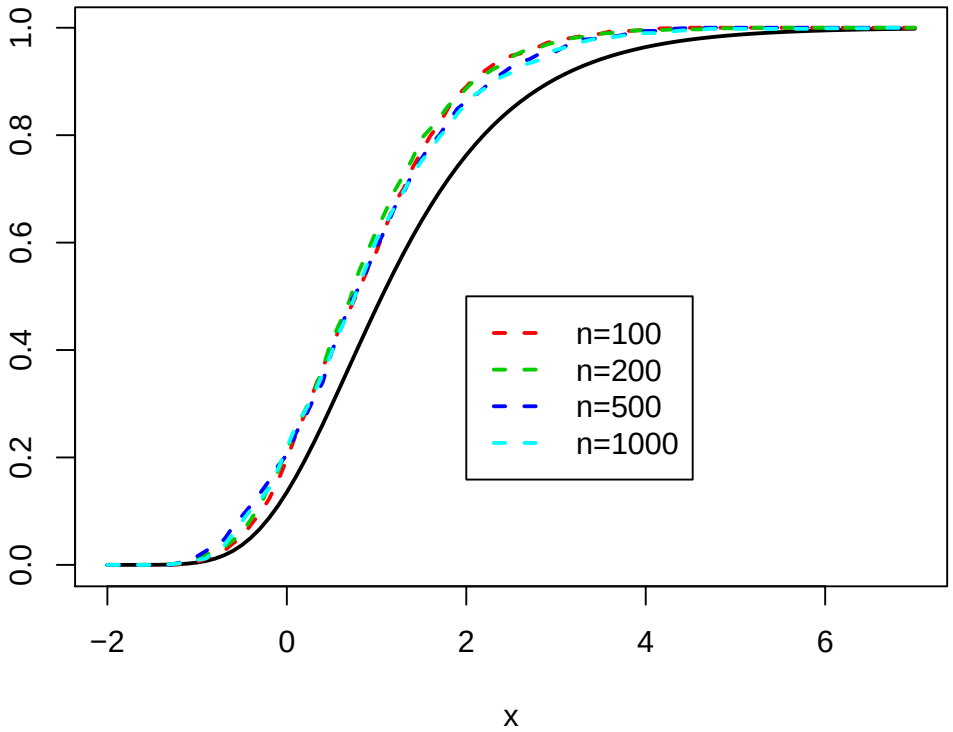
↪ pro BAR(1):

$$\text{logit}(\pi_t) = \begin{cases} \beta_0 + \beta_1 Y_{t-1}, & \text{t=1,...,m,} \\ \beta_0^* + \beta_1^* Y_{t-1}, & t = m+1, \ldots, n. \end{cases}$$

↪ pro BAR(p), $p > 1$, podobné

BAR(1) za H_0

$$t_n = \left(T_n - \sqrt{2\log\log(n)} - \frac{\log\log\log n}{2\sqrt{2\log\log(n)}}\right)\sqrt{2\log\log n} + \frac{1}{2}\log(\pi)$$



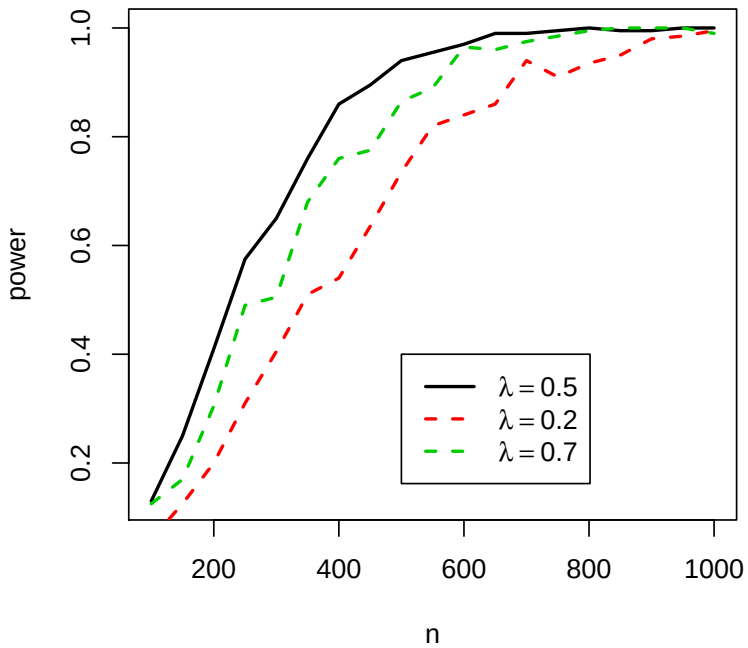
↪ volba $\beta_0 = 2$, $\beta_1 = -2$

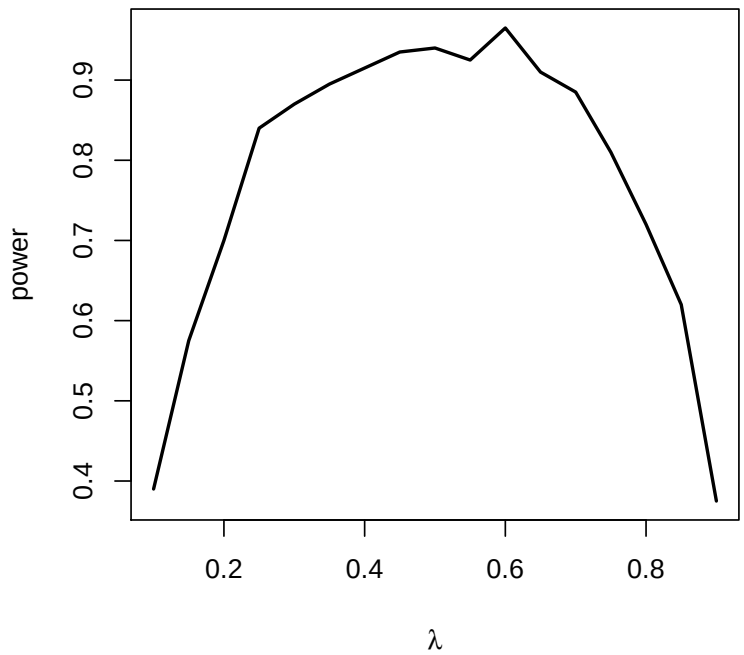
↪ hladina testu $< 0.05 \rightsquigarrow$ konzervativní test

BAR(1) za H_1

- data délky n, bod změny $m = \lambda \cdot n$
- $\beta_0 = 2$, $\beta_1 = -2 \rightsquigarrow$ změna na $\beta_0^* = \beta_0 + \delta_0$

$$\delta_0 = -1$$

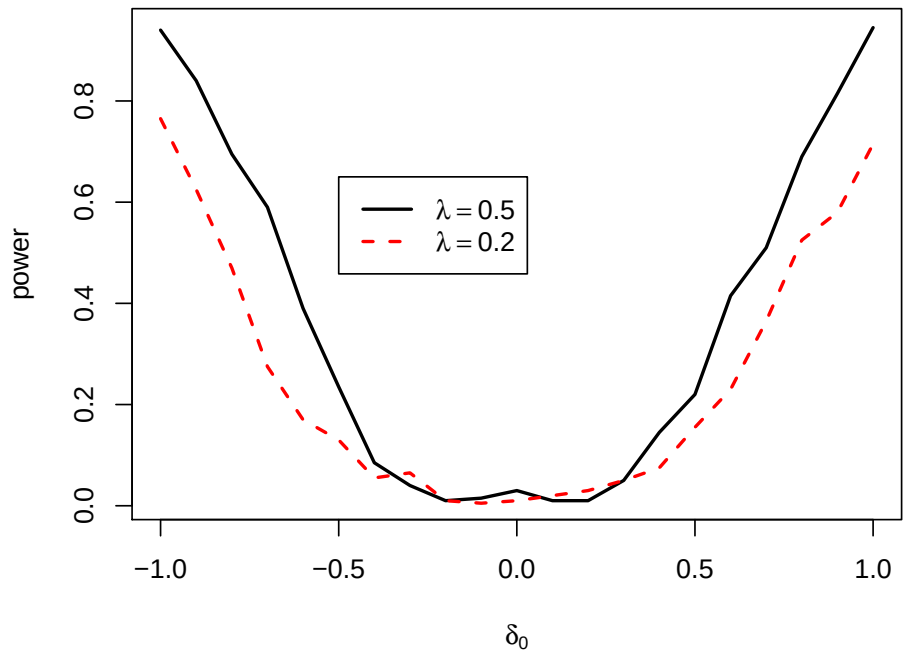




$$(p_{01}, p_{11}): \quad (0.88, 0.50) \rightarrow (0.73, 0.27)$$

BAR(1) za H_1

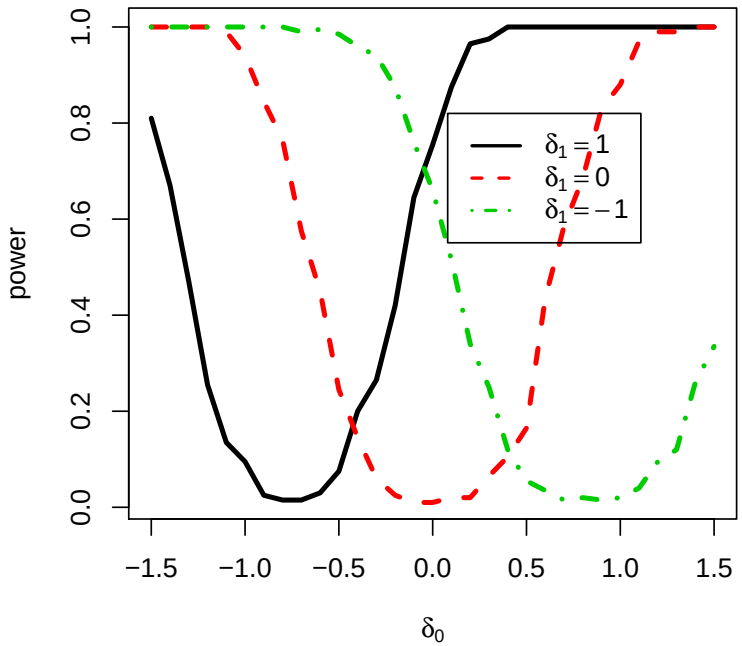
- data délky $n = 500$, bod změny $m = \lambda \cdot n$
- $\beta_0 = 2$, $\beta_1 = -2 \rightsquigarrow$ změna na $\beta_0^* = \beta_0 + \delta_0$

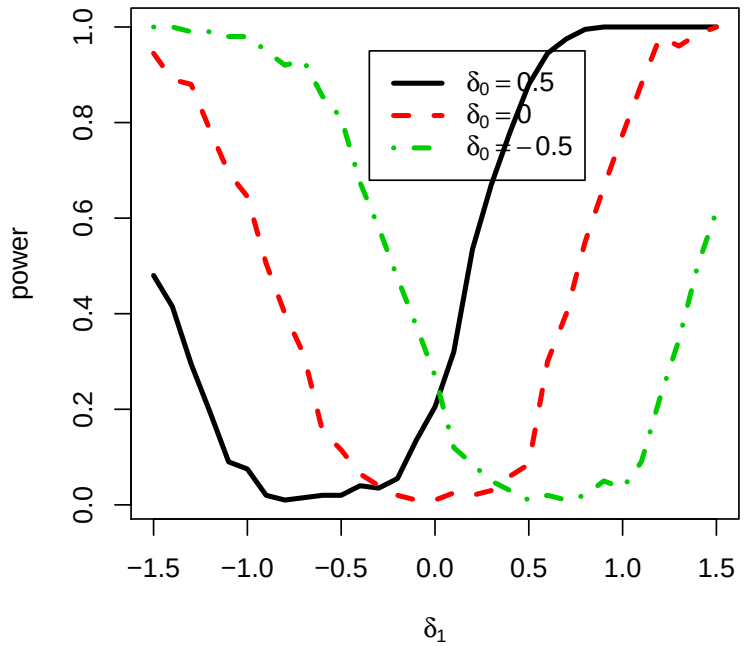


$$\delta_0 = 0.75 \leftrightsquigarrow (p_{01}, p_{11}): \quad (0.88, 0.50) \rightarrow (0.78, 0.32)$$

BAR(1) se změnou v obou parametrech

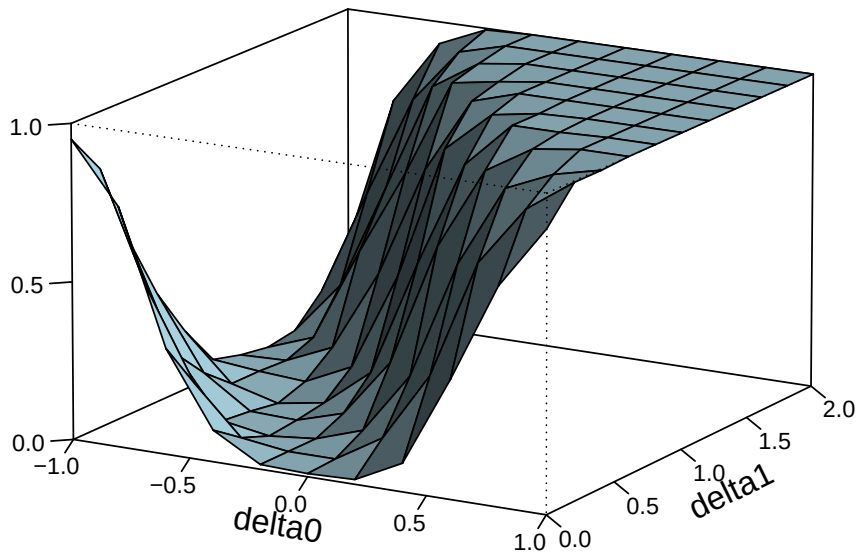
- data délky $n = 500$, bod změny $m = 1/2 \cdot n$
- $\beta_0 = 2$, $\beta_1 = -2 \rightsquigarrow$ změna na $\beta_0^* = \beta_0 + \delta_0$





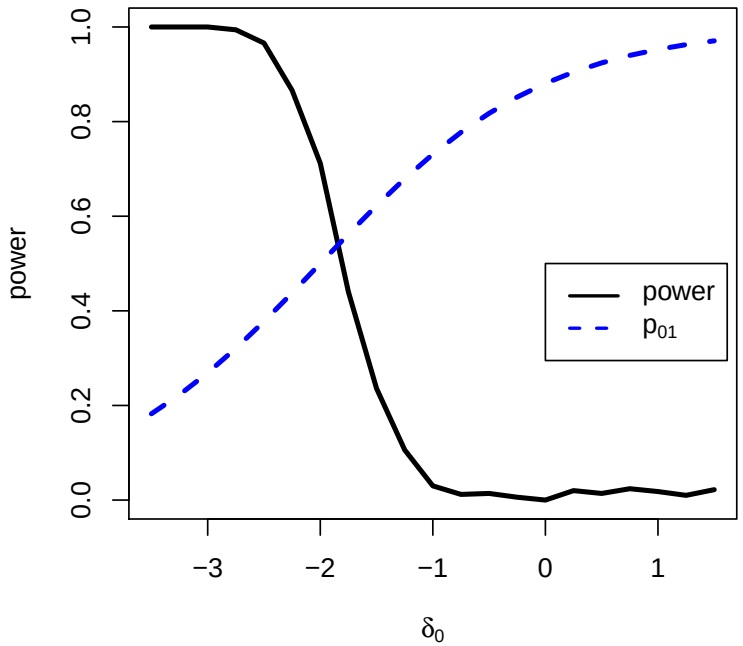
BAR(1) se změnou v obou parametrech

- data délky $n = 500$, bod změny $m = 1/2 \cdot n$
- $\beta_0 = 2$, $\beta_1 = -2 \rightsquigarrow$ změna na $\beta_0^* = \beta_0 + \delta_0$



BAR(1) se změnou v obou parametrech

- data délky $n = 500$, bod změny $m = 1/2 \cdot n$
- $\beta_0 = 2$, $\beta_1 = -2 \rightsquigarrow$ změna na $\beta_0^* = \beta_0 + \delta_0$



Obsah

Výskyt srážek

- denní výskyt srážek
- 12 modelů pro každý měsíc zvlášť
- zahrnut vliv teploty T_t

$$\text{logit}(\pi_t) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 T_t$$

- test, zda došlo v modelu ke změně

Výskyt srážek

- denní výskyt srážek
- 12 modelů pro každý měsíc zvlášť
- zahrnut vliv teploty T_t

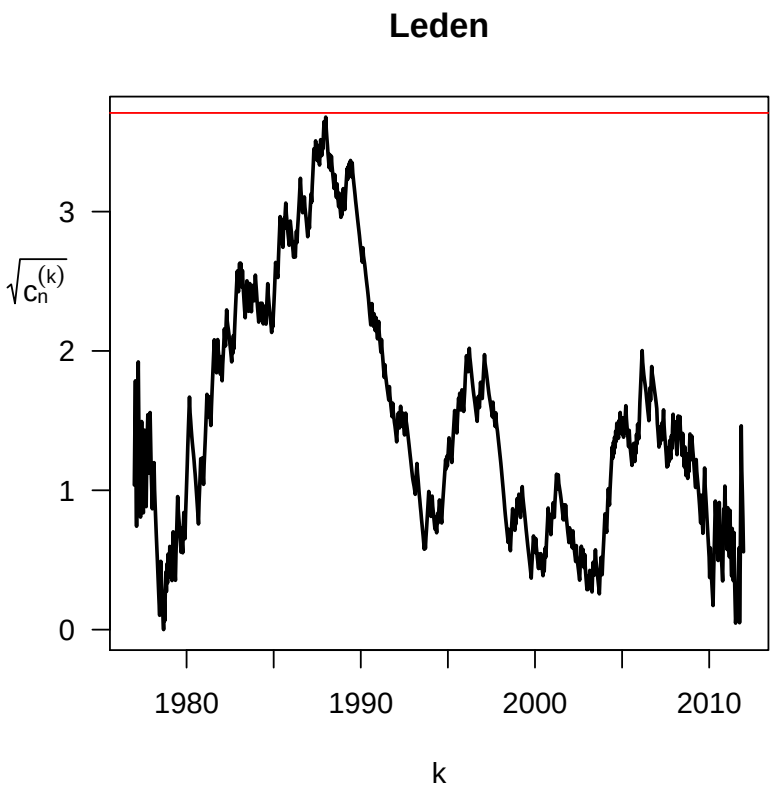
$$\text{logit}(\pi_t) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 T_t$$

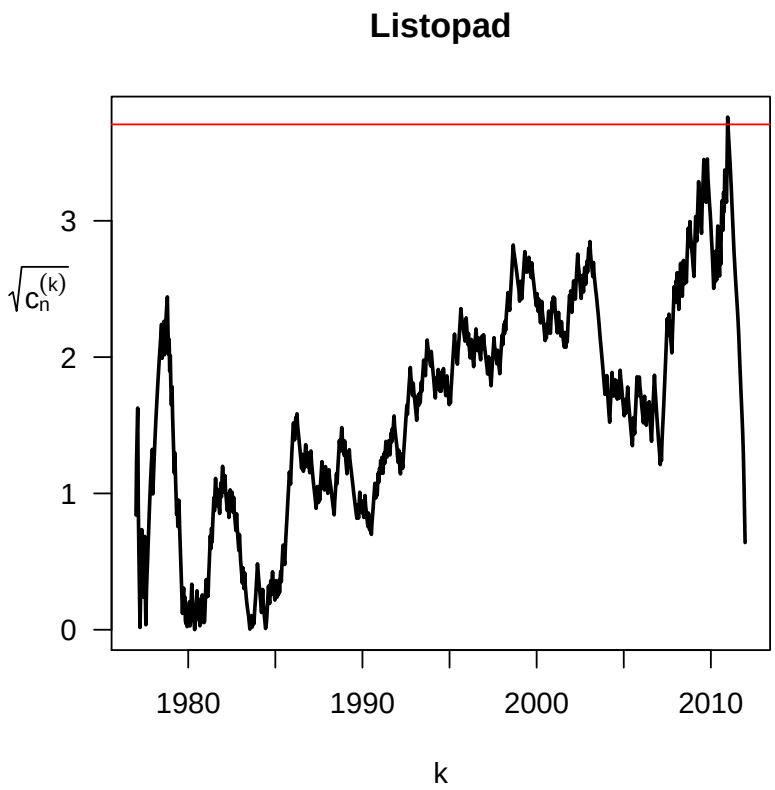
- test, zda došlo v modelu ke změně

Výsledek:

↪ listopad ($p = 0.0432$), leden ($p = 0.0529$)

↪ ostatní nevýznamné


Leden
$\sqrt{c_n^{(k)}}$
3
2
1
0
1980
1990
2000
2010
k

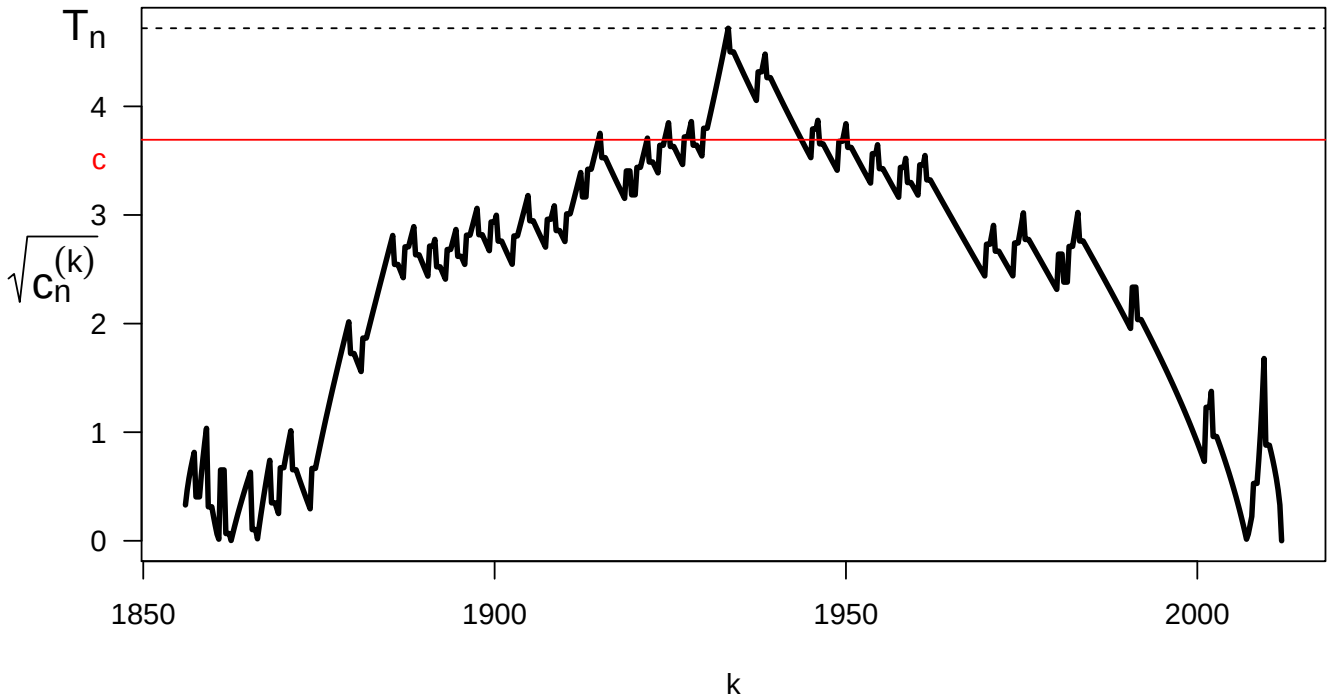

Listopad
$\sqrt{c_n^{(k)}}$
3
2
1
0
1980
1990
2000
2010
k

Recese USA

- čtvrtletní ukazatele recese v USA z let 1855–2011
- BAR(3) model $\rightsquigarrow$ test a detekce změny

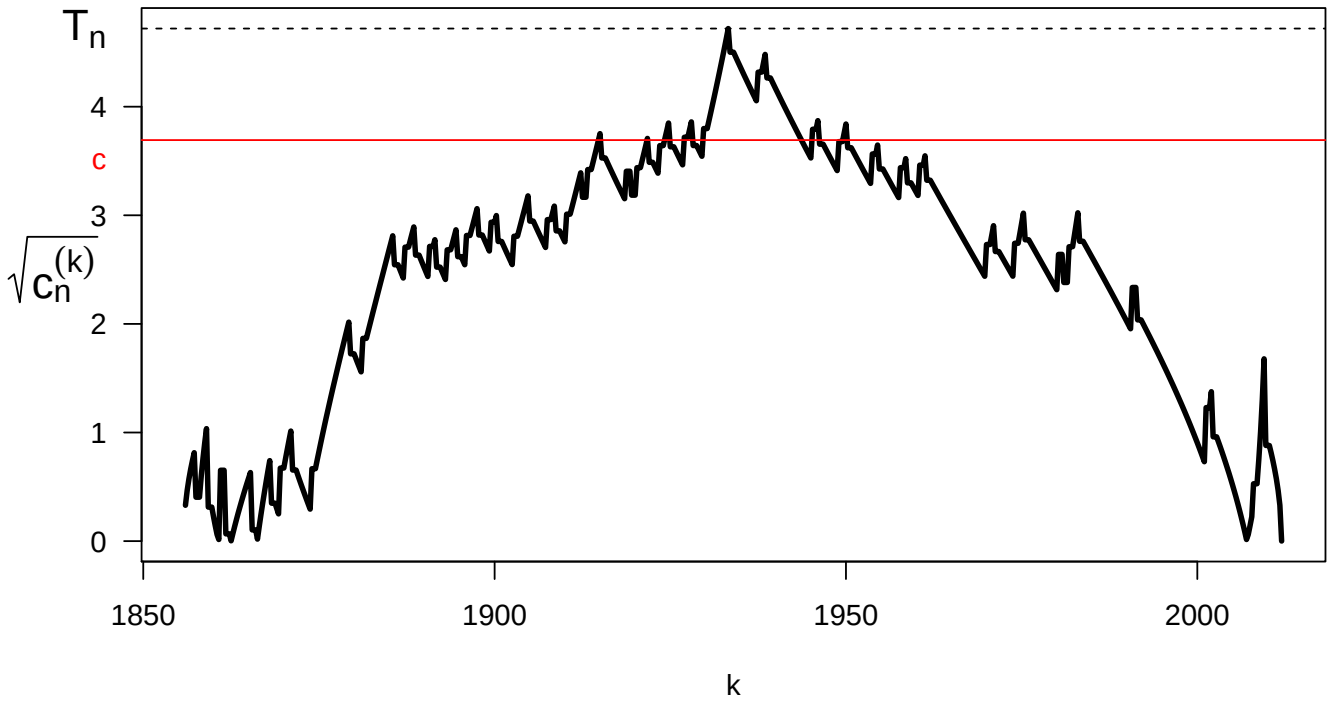
Recese USA

- čtvrtletní ukazatele recese v USA z let 1855–2011
- BAR(3) model $\rightsquigarrow$ test a detekce změny



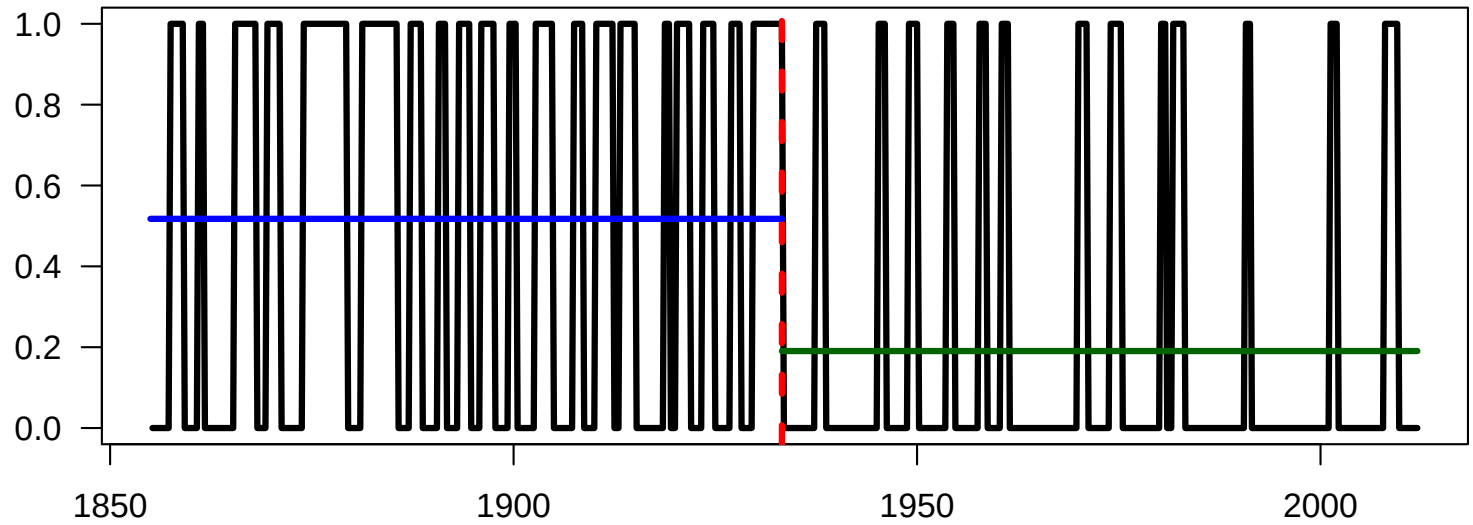
- maximum $T_n = 4.7$, kritická hodnota $c_{0.05} = 3.7$
- zamítáme H_0 s asymptotickou p-hodnotou 0.007

Recese USA II.



- maximum dosaženo pro první čtvrtletí 1933
- kritická hodnota překročena v celém období 1927–1946
- výsledky jsou v souladu s očekáváním, že ke změně došlo v období kolem 2. světové války

Recese USA III.



Výsledný odhadnutý model:

$$\text{logit}[\widehat{\pi}_t] = \begin{cases} -1.52 + 21.16 y_{t-1} - 2.96 \cdot 10^{-9} y_{t-2} - 17.99 y_{t-3}, & t = 4, \ldots, 313, \\ -2.80 + 21.54 y_{t-1} + 1.32 \cdot 10^{-8} y_{t-2} - 1.834 y_{t-3}, & t = 314, \ldots, 628. \end{cases}$$

Závěr

Shrnutí

- BAR modely pro binární časové řady
- testování změny

Možné další kroky

- odvození testu proti obecnější alternativě (změna ve více parametrech)
- využití jiných přístupů (bootstrap,...)
- navržení podobné procedury pro další GLM časové řady (Poisson, ...)

Děkuji za pozornost!

Literatura

- Antoch, J., Gregoire, G., and Jarůšková, D. (2004). Detection of structural changes in generalized linear models. *Stat. and Probab. Lett.*, 69:315–332.
- Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* Wiley, New York.
- Einmahl, U. and Mason, D. M. (1989). Darlin-Erdös theorems for martingales. *J. Theoret. Probab.*, 2(4), 437–460.
- Hall, P. and Heyde, C. C. (1980). *Martingale limit theory and its application.* Probability and mathematical statistics. Academic Press.
- Kedem, B. and Fokianos, K. (2002). *Regression Models for Time Series Analysis.* Wiley, New York.
- Philipp, W. and Stout, W. (1986). Invariance principles for martingales and sums of independent random variables. *Mathematische Zeitschrift*, 192:253–264.

Data recese USA:

- De Lucca, G. and Carfora, A. (2011). Predicting binary time series with a long memory structure: an application to U.S. recession time series. Available at http://www.smye2011.org/.
- Kauppi, H. and Saikkonen, P. (2008). Predicting U.S. recessions with dynamic binary response models. *Rev. Econ. Stat.*, 90(4):777–791.
- Startz, R. (2008). Binomial autoregressive moving average models with an application to U.S. recession. *J. Bus. Econom. Statist.*, 26:1–8.