

Testy dobré shody pro časové řady s diskrétními veličinami

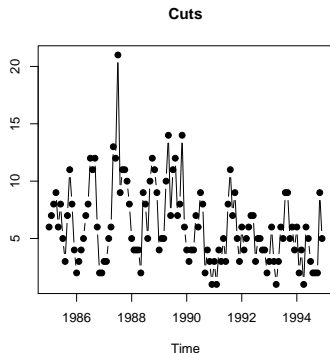
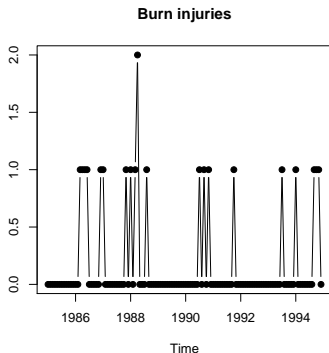
Šárka Hudecová, Marie Hušková a Simos G. Meintanis

KPMS MFF UK

Robust 2016

Motivace: Časová řada diskretních veličin

- $\{Y_t\}_{t \in \mathbb{N}}$ časová řada nezáporných celočíselných veličin
- klasické ARMA modely **nelze využít**



- speciální modely: INAR, INARCH, ...
- testy dobré shody (GOF)

- 1 **Modely INAR a INARCH**
- 2 **Testová statistika**
- 3 **Asymptotické rozdělení**
- 4 **Simulace**
- 5 **Příklad**

INAR(1) model

- klasická autoregrese: $Y_t = \phi \cdot Y_{t-1} + \varepsilon_t$
- INAR(1)

$$Y_t = p \circ Y_{t-1} + \varepsilon_t,$$

kde ε_t jsou iid nezáporné, diskrétní, nezávislé na Y_{t-1} ,
 $\text{var } \varepsilon_t < \infty$ a \circ je operátor (Steutel-van Harn)

$$p \circ Y_{t-1} = \sum_{i=1}^{Y_{t-1}} U_{t,i},$$

kde $U_{t,i}$ jsou $\text{Alt}(p)$ nezávislé na Y_{t-1} a ε_t ,

- $\hookrightarrow p \in (0, 1) \rightsquigarrow$ stacionární, ergodický a absolutně regulární
- \hookrightarrow korelace $\text{cor}(Y_t, Y_{t-k}) = p^k$
- \hookrightarrow rozdělení $\varepsilon_t \iff$ rozdělení Y_t
 Poissonovo (speciální postavení), negativně binomické,
 směs Poissonových rozdělení aj. \iff **testy dobré shody**

INARCH(1) model

- $\mathcal{I}_t = \sigma\{Y_s, s \leq t\}$ informace do času t
- INARCH(1)

$$Y_t | \mathcal{I}_{t-1} \sim \text{Po}(\lambda_t), \quad \lambda_t = r(Y_{t-1}, \theta)$$

- Poissonovská lineární autoregrese $\iff \lambda_t = \theta_1 + Y_{t-1}\theta_2$
 - \hookrightarrow pro $\theta_1 > 0, \theta_2 \in (0, 1) \rightsquigarrow$ stacionární, ergodický
 - \hookrightarrow korelace $\text{cor}(Y_t, Y_{t-k}) = \theta_2^k$
 - $\hookrightarrow \text{var } Y_t > \text{E} Y_t$ (overdispersion)
- negativně binomický INARCH ...

Test dobré shody s modelem

- obecný tvar modelu

$$Y_t | \mathcal{I}_{t-1} \sim F(\cdot; \mathcal{I}_{t-1}, \theta)$$

data Y_1, \dots, Y_T

- empirická pravděpodobnostní vytvořující funkce (PGF)
 - X diskrétní $\rightsquigarrow g_X(u) = \mathbb{E}u^X$, $u \in [0, 1]$,
 - pro Y_1, \dots, Y_T neparametrický odhad

$$\hat{g}_{T,Y}(u) = \frac{1}{T} \sum_{t=1}^T u^{Y_t}$$

- **idea:** porovnáme $\hat{g}_{T,Y}(u)$ se (semiparametrickým) odhadem využívajícím strukturu modelu

Testová statistika pro INAR model

$H_0 : \{Y_t\}$ se řídí modelem

$$Y_t = p \circ Y_{t-1} + \varepsilon_t,$$

pro $p \in (0, 1)$ a $g_\varepsilon \in \mathcal{G}_\theta = \{g_\varepsilon(\cdot, \theta); \theta \in \Theta\}$

- vytvěřující funkce:

$$\begin{aligned} g_Y(u) &= E u^{Y_t} = E E[u^{Y_t} | Y_{t-1}] = E E[u^{\sum_{i=1}^{Y_{t-1}} U_{t,i} + \varepsilon_t} | Y_{t-1}] = E[(1 + p(u - 1))^{Y_{t-1}}] \cdot g_\varepsilon(u, \theta) \\ &= g_Y(1 + p(u - 1)) \cdot g_\varepsilon(u, \theta) \end{aligned}$$

- semiparametrický odhad

$$\hat{g}_{T0}(u) = \hat{g}_{T,Y}(1 + \hat{p}(u - 1)) \cdot g_\varepsilon(u, \hat{\theta}),$$

kde \hat{p} a $\hat{\theta}$ jsou vhodné odhady a $\hat{g}_{T,Y}$ je neparametrický odhad g_Y

- testová statistika

$$S_T = T \int_0^1 (\hat{g}_{T,Y}(u) - \hat{g}_{T0}(u))^2 w(u) du,$$

kde $w(u) \geq 0$ je váhová funkce

Testová statistika pro INARCH model

$H_0 : \{Y_t\}$ se řídí modelem

$$Y_t | \mathcal{I}_{t-1} \sim \text{Po}(\lambda_t), \quad \lambda_t = \theta_1 + \theta_2 Y_{t-1}$$

pro $(\theta_1, \theta_2) \in \Theta = \{(\theta_1, \theta_2) : \theta_1 > 0, \theta_2 \in (0, 1)\}$

- vytvořující funkce

$$g_Y(u) = E u^{Y_t} = E E[u^{Y_t} | Y_{t-1}] = E[e^{(u-1)(\theta_1 + \theta_2 Y_{t-1})}] = e^{\theta_1(u-1)} g_Y(e^{(u-1)\theta_2})$$

- semiparametrický odhad

$$\widehat{g}_{T0}(u) = e^{\widehat{\theta}_1(u-1)} \widehat{g}_{T,Y}(e^{\widehat{\theta}_2(u-1)}),$$

kde $(\widehat{\theta}_1, \widehat{\theta}_2)$ je vhodný odhad

- testová statistika

$$S_T = T \int_0^1 (\widehat{g}_{T,Y}(u) - \widehat{g}_{T0}(u))^2 w(u) du$$

Předpoklady pro INAR

(A1) $\{Y_t\}$ se řídí INAR(1) za platnosti H_0

(A2) existuje $\partial g_\varepsilon / \partial \theta$ splňující pro $u \in [0, 1]$ a $|\theta - \theta_0| \leq D_2$

$$\left| \frac{\partial g_\varepsilon}{\partial \theta}(u; \theta) - \frac{\partial g_\varepsilon}{\partial \theta}(u; \theta_0) \right| \leq D_1 |\theta_0 - \theta| v(u) \quad \text{a} \quad \left| \frac{\partial g_\varepsilon}{\partial \theta}(u; \theta) \right| \leq D_3 v(u)$$

pro $D_1, D_2, D_3 > 0$ a vhodnou funkci $v(\cdot)$.

(A3) $0 < \int_0^1 w(u) du < \infty$, $\int_0^1 w(u) v^2(u) du < \infty$

(A4) odhad $\hat{\vartheta}_T = (\hat{\rho}_T, \hat{\theta}_T)'$ splňuje

$$\sqrt{T}(\hat{\vartheta}_T - \vartheta_0) = \frac{1}{\sqrt{T}} \sum_{t=1+q}^T \ell(\mathbf{Y}_{t-q}; \vartheta_0) + o_P(1),$$

kde $\mathbf{Y}_{t-q} = (Y_t, \dots, Y_{t-q})'$, $q \geq 1$ je pevné a $\ell = (\ell_1, \ell_2)'$, kde $\ell_j(\mathbf{Y}_{t-q}; \vartheta_0)$, $j = 1, 2$ jsou posloupnosti martingalových diferencí s konečným rozptylem

Asymptotické rozdělení pro INAR

Věta

Platí-li podmínky (A1)–(A4), pak S_T konverguje v distribuci k

$$\int_0^1 Z^2(u; p_0, \theta_0) w(u) du,$$

kde $\{Z(u; p_0, \theta_0), u \in [0, 1]\}$ je centrovaný Gaussovský proces s kovarianční strukturou $\mathbb{E}[Z_{q+1}(u_1; p_0, \theta_0) Z_{q+1}(u_2; p_0, \theta_0)]$ pro $u_1, u_2 \in [0, 1]$, kde

$$\begin{aligned} Z_t(u; p, \theta) = & u^{Y_t} - (1 + p(u - 1))^{Y_{t-1}} g_\varepsilon(u; \theta) + h_1(u; p, \theta) \ell_1(\mathbf{Y}_{t-q}; p, \theta) \\ & + h_2(u; p, \theta) \ell_2(\mathbf{Y}_{t-q}; p, \theta) \end{aligned}$$

a

$$h_1(u; p, \theta) = \frac{\partial g_Y(1 + p(u - 1))}{\partial p} g_\varepsilon(u; \theta), \quad h_2(u; p, \theta) = g_Y(1 + p(u - 1)) \frac{\partial g_\varepsilon(u; \theta)}{\partial \theta}.$$

Poznámky

- limitní rozdělení závisí na neznámých hodnotách
 - ↪ konzistentní odhad kovarianční struktury
 - ↪ parametrický bootstrap
- odhad $(\hat{\rho}, \hat{\theta})$ např. metodou nejmenších čtverců (CLS) \iff platí (A4)
- váhová funkce $w(u) = u^a, a \geq 0$

Rozdělení za alternativy

Alternativa $g_\varepsilon \notin G_\theta$

Theorem

Nechť se $\{Y_t\}$ řídí INAR(1) modelem a necht' platí (A3),

$(\hat{p}_T, \hat{\theta}_T)' \xrightarrow{P} (p, \theta_A)'$ a $g_\varepsilon(u; \theta)$ je spojitá v θ pro všechna $u \in [0, 1]$. Pak pro $T \rightarrow \infty$

$$\frac{S_T}{T} \xrightarrow{P} \int_0^1 \left[g_Y(1 + p(u-1))(g_\varepsilon(u) - g_\varepsilon(u; \theta_A)) \right]^2 w(u) du.$$

↪ konzistentní test

↪ alternativy porušení INAR struktury

Předpoklady pro INARCH

(B1) $\{Y_t\}$ se řídí INARCH(1) za platnosti H_0

(B2) odhady $\hat{\theta}_{1T}, \hat{\theta}_{2T}$ splňují

$$\sqrt{T}(\hat{\theta}_{jT} - \theta_{j0}) = \frac{1}{\sqrt{T}} \sum_{t=q+1}^T \ell_j(\mathbf{Y}_{t-q}; \theta_{10}, \theta_{20}) + o_P(1), \quad j = 1, 2,$$

kde $q \geq 1$ je pevné a $\ell_j(\mathbf{Y}_{t-q}; \theta_{10}, \theta_{20})$, $j = 1, 2$ jsou posloupnosti martingalových diferencí s konečným rozptylem

(B3) $0 < \int_0^1 w(u) du < \infty$

Asymptotické rozdělení pro INARCH

Theorem

Platí-li (B1)–(B3), pak S_T konverguje v distribuci k

$$\int_0^1 \mathcal{V}^2(u; \theta_{10}, \theta_{20}) w(u) du,$$

kde $\{\mathcal{V}(u; \theta_{10}, \theta_{20}), u \in [0, 1]\}$ je centrovaný Gaussovský proces s kovarianční strukturou $\mathbb{E}[\mathcal{V}_{q+1}(u_1; \theta_{10}, \theta_{20}) \mathcal{V}_{q+1}(u_2; \theta_{10}, \theta_{20})]$ pro $u_1, u_2 \in [0, 1]$, kde

$$\begin{aligned} \mathcal{V}_t(u; \theta_1, \theta_2) = & u^{Y_t} - \exp\{(\theta_1 + \theta_2 Y_{t-1})(u-1)\} + r_1(u; \theta_1, \theta_2) \ell_1(\mathbf{Y}_{t-q}; \theta_1, \theta_2) \\ & + r_2(u; \theta_1, \theta_2) \ell_2(\mathbf{Y}_{t-q}; \theta_1, \theta_2) \end{aligned}$$

a

$$r_1(u; \theta_1, \theta_2) = \mathbb{E}\left[e^{(\theta_1 + \theta_2 Y_1)(u-1)}(u-1)\right], \quad r_2(u; \theta_1, \theta_2) = \mathbb{E}\left[Y_1 e^{(\theta_1 + \theta_2 Y_1)(u-1)}(u-1)\right].$$

↪ parametrický bootstrap

Simulace

- test založený na parametrickém bootstrapu
 - CLS odhady, $w(u) = u^a$ pro $a = 0, 1, 2, 5$
 - $B = 500$; empirická síla spočtená z 500 opakování
 - $T = 50, 100, 500$
- hladina testu
- H_0 : Poissonovský INAR(1) \leftrightarrow síla pro
 - $\hookrightarrow g_\varepsilon$ negativně binomické
 - $\hookrightarrow g_\varepsilon$ směs dvou Poissonových rozdělení
 - $\hookrightarrow g_\varepsilon$ směs Poissonova a Diracovy míry v 0
 - \hookrightarrow Poissonovský INARCH(1)
- H_0 : Poissonovský INARCH(1) \leftrightarrow síla pro
 - \hookrightarrow negativně binomický INARCH(1)
 - \hookrightarrow Poissonovský INAR(1)
 - \hookrightarrow Poissonovský INARCH(1,1)
 - \hookrightarrow Poissonovský INARCH(1) se změnou v hladině

Hladina testu

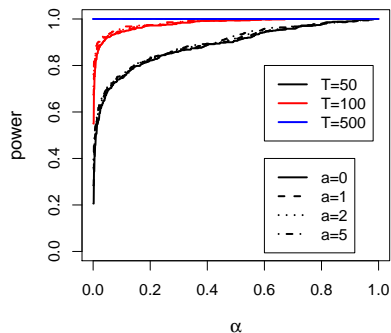
		$H_0 : \text{INAR}(1)$		
		α		
T	a	0.01	0.05	0.1
50	0	0.008	0.044	0.100
50	1	0.008	0.050	0.102
50	2	0.006	0.054	0.112
50	5	0.008	0.056	0.110
100	0	0.004	0.038	0.080
100	1	0.004	0.040	0.082
100	2	0.006	0.042	0.088
100	5	0.012	0.052	0.102
500	0	0.018	0.038	0.090
500	1	0.016	0.044	0.090
500	2	0.010	0.042	0.088
500	5	0.010	0.044	0.094

		$H_0 : \text{INARCH}(1)$		
		α		
T	a	0.01	0.05	0.1
50	0	0.006	0.038	0.082
50	1	0.008	0.036	0.092
50	2	0.006	0.040	0.090
50	5	0.004	0.044	0.090
100	0	0.002	0.036	0.094
100	1	0.004	0.042	0.096
100	2	0.002	0.042	0.096
100	5	0.006	0.044	0.094
500	0	0.006	0.036	0.080
500	1	0.006	0.034	0.092
500	2	0.004	0.042	0.092
500	5	0.008	0.040	0.102

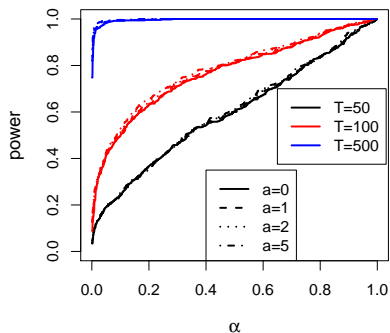
Síla testu: Negativně binomický INAR

g_ε je NB tak, že $\text{var } \varepsilon_t = \theta(1 + \theta/r)$

INAR H_1 :NB $r=2$



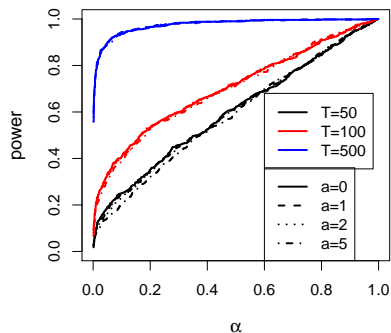
INAR H_1 :NB $r=5$



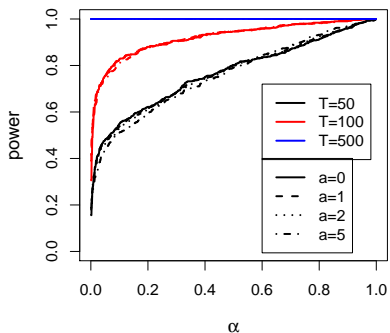
Síla testu: Směs Po a Diracovy míry

g_ε odpovídá $\psi D_0 + (1 - \psi) \text{Po}(\lambda)$

INAR H_1 : Dirac $\psi = 0.1$

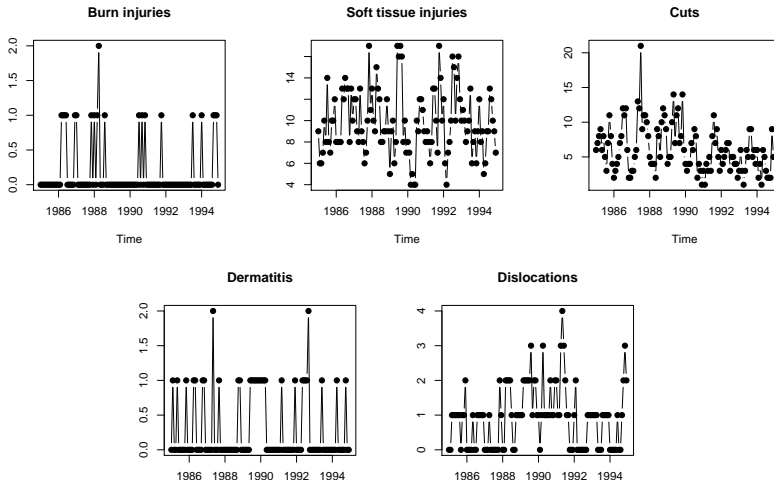


INAR H_1 : Dirac $\psi = 0.2$



Příklad

British Columbia Workers Compensation Board: měsíční počty žádostí o invalidní dávky, Freeland (1998)



Příklad — pokrač.

Lze tato data úspěšně popsat INAR(1) modelem?

Řada	<i>a</i>				
	0	1	2	5	10
1	0.653	0.652	0.698	0.771	0.862
2	0.561	0.596	0.579	0.662	0.665
3	0.020	0.002	0.001	0.000	0.000
4	0.270	0.270	0.295	0.386	0.506
5	0.444	0.486	0.579	0.661	0.749

- pro řady 1,2,4,5 se model INAR(1) jeví jako adekvátní
- pro řadu 3 nelze INAR(1) použít

Závěrečné poznámky

- snadné rozšíření pro modely vyšších řádů $\text{INAR}(p)$, $\text{INARCH}(p)$, $p > 1$ a na obecnější INARCH
- simulace: test má dostatečnou sílu pro různé odchylky od Poissonova rozdělení i pro změnu struktury modelu
- hledání „optimální“ váhové funkce

Děkuji za pozornost

Literatura

Příspěvek je založený na článku



Hudecová, Š., Hušková, M., Meintenis, S.G. (2015): Tests for time series of counts based on the probability-generating function. *Statistics* 49(2), 316-337.

Data



Freeland R.K. (1998): *Statistical Analysis of Discrete Time series with Application to the Analysis of Workers' Compensation Claims Data* (PhD thesis). Management Science Division, Faculty of Commerce and Business Administration, University of British Columbia.



Zhu F. and Joe H. (2006) Modelling count data time series with Markov processes based on binomial thinning. *J Time Series Anal.* 27: 725–738.

Modely pro diskrétní časové řady



McKenzie E. (2003): Discrete variate time series. In Shanbhag and Rao (eds.): *Handbook of Statistics* 21: 573–606.



Fokianos K. (2011): Some progress in count time series models. *Statistics* 45: 49–58.