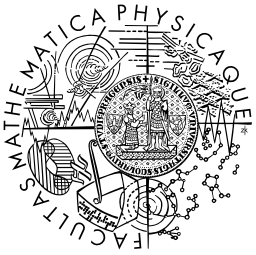


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Monika Kunayová

Modely binárních časových řad

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr.

Studijní program: Matematika (B1101)

Studijní obor: MFMAT (1103R024)

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Moja úprimná vďaka patrí vedúcej práce, RNDr. Jitke Zichovej, Dr., za usmernenia, pripomienky a odborné konzultácie pri písaní tejto práce.

Název práce: Modely binárních časových řad

Autor: Monika Kunayová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Jitka Zichová, Dr., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Táto práca sa venuje časových radom binárnych premenných, ktoré sa vyskytujú v mnohých spoločenských sférach. Indikátor môže značiť prekročenie určitej hodnoty alebo výskyt nejakého javu. V práci sa zaoberáme logistickým rozdelením a jeho vlastnosťami, parciálnou vierohodnostnou funkciou, ktorá umožňuje pracovať so závislými dátami, a odvodíme užitočné vzťahy pre praktickú aplikáciu, ktorá pozostáva zo simulácie časového radu a z analýzy reálnych dát pomocou voľne šíriteľného softvéru R.

Klíčová slova: binárne časové rady, logistická regresia, metóda maximálnej vierohodnosti

Title: Models of binary time series

Author: Monika Kunayová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jitka Zichová, Dr., Department of Probability and Mathematical Statistics

Abstract: This bachelor thesis deals with the time series of binary variables that exist in many social spheres. The indicator may denote a certain value being exceeded or a phenomenon occurring. We study a model of logistic autoregression and its properties, partial likelihood function which allows us to work with dependent data, and derive useful relationships for a practical application that consists of time series simulation and real data analysis using free software R.

Keywords: binary time series, logistic regression, maximum likelihood estimation

Obsah

Úvod

Časové rady možno chápať ako postupnosti hodnôt chronologicky zaznamenaných v čase. Druhý pohľad, ktorý nám umožňuje modelovanie na základe teórie pravdepodobnosti, je pozerať sa na časový rad ako špeciálny prípad náhodných procesov (Cipra, 2008). Historicky boli skúmané už celé storočia a postupne nadobúdajú na význame, keďže ich môžeme nájsť takmer v každej sfére života, či už je to ekonometria, financie, technika, prírodné vedy, astronómia alebo demografia. Modelovanie a predpovedanie budúcich hodnôt časových radov je dôležitou časťou dátovej analýzy.

Cieľom analýzy časových radov je podľa knihy (Cipra, 2008) pochopiť mechanizmus, ktorý generuje analyzované dáta, testovať hypotézy, predpovedať budúci vývoj, riadiť a optimalizovať. Analýza berie do úvahy vnútornú štruktúru dát a pozorovateľné trendy v nich. Predikcia je nevyhnutná v ekonomických modeloch, kedy sa podnik rozhoduje o budúcej výrobe alebo investor obchodujúci s menami či akciami o kúpe alebo predaji, pri predpovediach počasia, v medicíne môže byť dôležitý počet chorých kvôli zabezpečeniu dostatočného počtu vakcín alebo aj v konaní jednotlivca pri kúpe auta či domu.

V praktických aplikáciách sa často stretávame s potrebou predikcie budúcej hodnoty binárnej premennej, ktorá môže nadobúdať len dve hodnoty – nastane, resp. nenastane udalosť. Na základe aj už spomínaných príkladov časových radov možno odvodiť binárne časové rady. Napríklad bude/nebude prekročená určitá hranica ceny akcie na burze, bude/nebude prekročená typická hodnota úhrnu zrážok, bude/nebude prekročený kritický počet infikovaných ľudí smrteľnou chorobou, teda hrozí/nehrozí epidémia.

Jednou z najčastejšie využívaných metód analýzy časových radov je regresná analýza, ktorá sa zaoberá vzťahom medzi závislou (výstupnou) premennou a jednou alebo viacerými nezávislými (vysvetľujúcimi) premennými. Základnou metódou regresnej analýzy je lineárna regresia, ktorá však vyžaduje spojitosť, normalitu a nezávislosť dát. V súčasnosti na význame naberajú tzv. zovšeobecnené lineárne modely, ktoré za určitých podmienok možno použiť aj v prípade kategoriálnych a spočítateľných dát (Kedem a Fokianos, 2002). V prípade binárnych časových radov tento problém rieši logistická regresia, v ktorej závislá premenná nadobúda iba hodnoty 0/1, úspech/neúspech.

Cieľom bakalárskej práce je zoznámenie sa s vybranými prístupmi k modelovaniu binárnych časových radov, ich podrobný popis a odvodenie vzorcov, ilustrácia odhadových procedúr na simulovaných dátach a analýza reálneho časového radu nula-jednotkovej povahy. Práca je rozdelená na teoretickú a praktickú časť. V prvej kapitole sa budeme zaoberať teoretickými vzťahmi, ktoré sú základom modelu logistickej autoregresie. Praktická časť práce pozostáva z dvoch kapitol. V prvej z nich využijeme naštudované poznatky pri simulácii procesu a odhadovaní jeho parametrov. Poslednou kapitolou je analýza reálneho binárneho časového radu, kde nájdeme aj predikciu hodnoty na nasledujúce obdobie.

1. Logistická regresia v časových radoch

1.1 Logistické rozdelenie

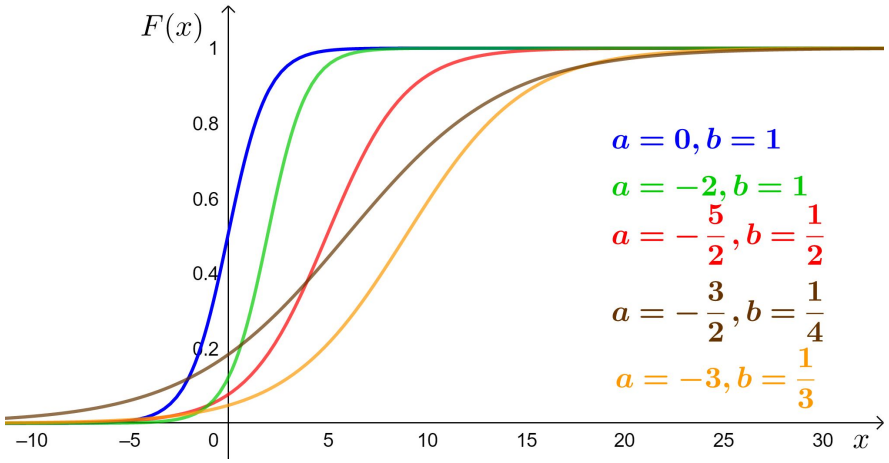
Jedným z významných pravdepodobnostných rozdelení je logistické rozdelenie. Jeho dôležitosť spočíva v širokých praktických aplikáciách. Používa sa napríklad na odhadovanie rastu ľudskej populácie a iných demografických dát, poľnohospodárskej produkcie a dokonca aj v biológii. Pre naše účely najdôležitejšie je využitie v logistickej regresii, pri modelovaní a predikovaní časových radov, ktoré môžu nadobúdať len dve hodnoty : 1=úspech, 0=neúspech.

Definícia 1 (Anděl, 2007, str. 23)**.** *Nech $a \in \mathbb{R}$, $b > 0$ sú dané čísla. Distribučná funkcia* logistického rozdelenia *je*

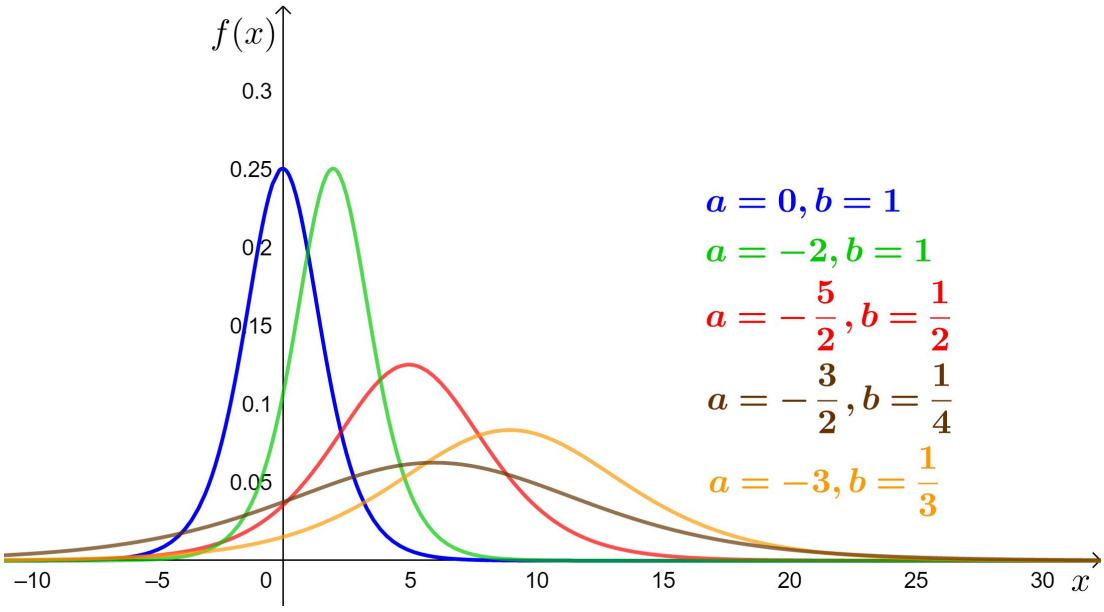
$$F(x) = \frac{1}{1 + e^{-(a+bx)}}, \quad x \in \mathbb{R} \tag{1.1}$$

Z distribučnej funkcie vieme pomocou derivácie určiť hustotu rozdelenia:

$$f(x) = \frac{dF}{dx} = \frac{b \cdot e^{-(a+bx)}}{(1 + e^{-(a+bx)})^2}, \quad x \in \mathbb{R} \tag{1.2}$$



Obrázok 1.1: Distribučná funkcia logistického rozdelenia pre rôzne hodnoty parametrov a, b.



Obrázok 1.2: Hustota logistického rozdelenia pre rôzne hodnoty parametrov a, b.

Odvodíme strednú hodnotu logistického rozdelenia

$$
\begin{aligned}
EX &= \int_{-\infty}^{\infty} x \cdot \frac{b \cdot e^{-(a+bx)}}{(1+e^{-(a+bx)})^2}\mathrm{d}x = \left\{ \begin{array}{r} a+bx = t, \quad t \in (-\infty,\infty) \\ b \cdot \mathrm{d}x = \mathrm{d}t \\ x = \dfrac{t-a}{b} \end{array} \right\} = \\
&= \int_{-\infty}^{\infty} \frac{t-a}{b} \cdot \frac{e^{-t}}{(1+e^{-t})^2}\mathrm{d}t = \frac{1}{b} \cdot \int_{-\infty}^{\infty} \frac{t \cdot e^{-t}}{(1+e^{-t})^2}\mathrm{d}t - \frac{a}{b}\int_{-\infty}^{\infty} \frac{e^{-t}}{(1+e^{-t})^2}\mathrm{d}t = \\
&= \left\{ \begin{array}{r} \dfrac{1}{1+e^{-t}} = y, \quad y \in (0,1) \\ \dfrac{e^{-t}}{(1+e^{-t})^2}\mathrm{d}t = \mathrm{d}y \\ \dfrac{1-y}{y} = e^{-t} \\ \ln\dfrac{1-y}{y} = -t \\ t = \ln\dfrac{y}{1-y} \end{array} \right\} = \\
&= \frac{1}{b}\int_0^1 \ln y \mathrm{d}y - \frac{1}{b}\int_0^1 \ln(1-y)\mathrm{d}y - \frac{a}{b}\left[y\right]_0^1 = \\
&= -\frac{a}{b}.
\end{aligned}
\tag{1.3}
$$

Posledná rovnosť vyplýva z jednoduchej substitúcie v druhom integrále

$$
\begin{aligned}
1-y &= u, \quad u \in (0,1) \\
-\mathrm{d}y &= \mathrm{d}u.
\end{aligned}
$$

Potom

$$\frac{1}{b}\int_0^1 \ln y \mathrm{d}y - \frac{1}{b}\int_0^1 \ln(1-y)\mathrm{d}y = 0.$$

Ku výpočtu rozptylu logistického rozdelenia potrebujeme vypočítať aj druhý moment

$$EX^2 = \int_{-\infty}^{\infty} x^2 \cdot \frac{b\cdot e^{-(a+bx)}}{(1+e^{-(a+bx)})^2}\mathrm{d}x = \left\{ \begin{array}{c} a+bx=t, \quad t\in(-\infty,\infty) \\ b\cdot \mathrm{d}x = \mathrm{d}t \\ x = \dfrac{t-a}{b} \end{array} \right\} =$$

$$= \int_{-\infty}^{\infty} \frac{(t-a)^2}{b^2}\cdot\frac{e^{-t}}{(1+e^{-t})^2}\mathrm{d}t =$$

$$= \frac{1}{b^2}\int_{-\infty}^{\infty}\frac{t^2e^{-t}}{(1+e^{-t})^2}\mathrm{d}t - \frac{2a}{b^2}\int_{-\infty}^{\infty}\frac{te^{-t}}{(1+e^{-t})^2}\mathrm{d}t + \frac{a^2}{b^2}\int_{-\infty}^{\infty}\frac{e^{-t}}{(1+e^{-t})^2}\mathrm{d}t.$$

Z výpočtu strednej hodnoty už vieme, že druhý integrál je nulový a tretí integrál je rovný 1. Potrebujeme teda dopočítať prvý integrál. Všimneme si, že v integrále sa nachádza párna funkcia a osobitne tento integrál vyriešime.

$$\int_{-\infty}^{\infty}\frac{t^2e^{-t}}{(1+e^{-t})^2}\mathrm{d}t = 2\int_0^{\infty}\frac{t^2e^{-t}}{(1+e^{-t})^2}\mathrm{d}t = 2\int_0^{\infty} t^2\sum_{n=1}^{\infty} n(-1)^{n-1}e^{-nt}\mathrm{d}t =$$

$$= 2\sum_{n=1}^{\infty} n(-1)^{n-1}\int_0^{\infty} t^2e^{-nt}\mathrm{d}t =$$

$$= 2\sum_{n=1}^{\infty} n(-1)^{n-1}\frac{2}{n^3} = 4\sum_{n=1}^{\infty}(-1)^{n-1}\frac{1}{n^2} = \frac{\pi^2}{3}.$$

Výraz $\frac{e^{-t}}{(1+e^{-t})^2}$ sme nahradili mocninným radom, vďaka čomu sme mohli zameniť poradie sumy a integrálu. Dokážeme si ešte, že výraz takémuto radu naozaj zodpovedá. Označme si $q := e^{-t}$.

$$\frac{1}{1+q} = \frac{1}{1-(-q)} = 1 + (-q) + (-q)^2 + .../()'$$

$$\frac{-1}{(1+q)^2} = 0 - 1 + 2q - 3q^2 + .../\cdot(-q)$$

$$\frac{q}{(1+q)^2} = q - 2q^2 + 3q^3 - 4q^4... = \sum_{n=1}^{\infty}(-1)^{n-1}nq^n = \sum_{n=1}^{\infty}(-1)^{n-1}ne^{-nt}.$$

Teraz už rozptyl dopočítame ľahko

$$VarX = EX^2 - (EX)^2 = \frac{1}{b^2}\frac{\pi^2}{3} + \frac{a^2}{b^2} - \left(\frac{-a}{b}\right)^2 = \frac{1}{b^2}\frac{\pi^2}{3}. \tag{1.4}$$

Štandardizované logistické rozdelenie je logistické rozdelenie, v ktorom sú $a = 0, b = 1$. Dosadením do definície 1 a 1.2 – 1.4 dostaneme

$$\mathrm{F}_l(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}, \tag{1.5}$$

$$f_l(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^x}{(1+e^x)^2}, \tag{1.6}$$

$$\begin{aligned} EX &= 0, \\ VarX &= \frac{\pi^2}{3}. \end{aligned} \tag{1.7}$$

1.2 Metóda maximálnej vierohodnosti

Metóda maximálnej vierohodnosti je jedna z najčastejšie používaných metód na hľadanie bodového odhadu neznámeho parametra $\boldsymbol{\theta}$ v určitom pravdepodobnostnom rozdelení. Odhad $\hat{\boldsymbol{\theta}}$ maximalizuje vierohodnostnú funkciu, prípadne jej logaritmus.

Definícia 2 (Kedem a Fokianos, 2002, str. 3)**.** *Nech $\mathcal{F}_t$, t=0,1,... je rastúca postupnosť σ-algebier, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$a nech Y_1,Y_2,... je postupnosť náhodných veličín na spoločnom pravdepodobnostnom priestore takom, že Y_t je $\mathcal{F}_t$-merateľná. Označme hustotu Y_t pri danom $\mathcal{F}_{t-1}$ ako $f_t(y_t;\boldsymbol{\theta})$, kde $\boldsymbol{\theta} \in \mathbb{R}^p$ je fixný parameter.* Parciálna vierohodnostná funkcia *vzhľadom ku $\boldsymbol{\theta}, \mathcal{F}_t$,a veličinám Y_1, Y_2,...,Y_N je daná súčinom*

$$PL(\boldsymbol{\theta}; y_1,...,y_N) = \prod_{t=1}^{N} f_t(y_t;\theta). \tag{1.8}$$

Parciálna vierohodnosť nám umožňuje pracovať aj s dátami, ktoré nie sú nezávislé. Označme $f(y_1,...,y_N;\boldsymbol{\theta})$ združenú hustotu náhodných veličín Y_1,Y_2,...,Y_N a podmienené hustoty píšme v tvare $f_t(y_t;\boldsymbol{\theta}) = f(y_t;\boldsymbol{\theta}|y_1,...,y_{t-1})$; $t = 2,...,N$. S využitím vzťahu medzi združenou a podmienenou hustotou môžeme písať

$$\begin{aligned} &f(y_1;\boldsymbol{\theta}) \cdot f(y_2;\boldsymbol{\theta}|y_1) \cdot ... \cdot f(y_N;\boldsymbol{\theta}|y_1,...,y_{N-1}) = \\ &= f(y_1;\boldsymbol{\theta}) \cdot \frac{f(y_2,y_1;\boldsymbol{\theta})}{f(y_1;\boldsymbol{\theta})} \cdot \frac{f(y_3,y_2,y_1;\boldsymbol{\theta})}{f(y_1,y_2;\boldsymbol{\theta})} \cdot ... \cdot \frac{f(y_1,...,y_N;\boldsymbol{\theta})}{f(y_1,y_2,...,y_{N-1})} = \\ &= f(y_1,...,y_N;\boldsymbol{\theta}). \end{aligned}$$

Máme teda

$$PL(\boldsymbol{\theta}; y_1,...,y_N) = f(y_1,...,y_N;\boldsymbol{\theta}) = f(y_1;\boldsymbol{\theta})) \prod_{t=2}^{N} f(y_t;\boldsymbol{\theta}|y_1,...,y_{t-1}). \tag{1.9}$$

Ak sú veličiny Y_1,...,Y_N nezávislé, $PL(\boldsymbol{\theta}; y_1,...,y_N) = \prod_{t=1}^{N} f(y_t;\boldsymbol{\theta})$ je súčin marginálnych hustôt jednotlivých veličín. Pre diskrétne rozdelené veličiny máme

$$PL(\boldsymbol{\theta}; y_1,...,y_N) = P(Y_1 = y_1;\boldsymbol{\theta}) \cdot \prod_{t=2}^{N} P(Y_t = y_t;\boldsymbol{\theta}|Y_1 = y_1,...,Y_{t-1} = y_{t-1}), \tag{1.10}$$

pri nezávislosti Y_1,...,Y_N dostávame

$$PL(\boldsymbol{\theta}; y_1,...,y_N) = \prod_{t=1}^{N} P(Y_t = y_t; \boldsymbol{\theta}).$$

Naďalej nech Y_t je časový rad nadobúdajúci hodnoty 0 alebo 1 a k-rozmerný stĺpcový vektor $\boldsymbol{Z}_{t-1}, t = 1,2,3,...$ reprezentuje množinu regresorov. Chceme odhadovať podmienenú pravdepodobnosť úspechu

$$\pi_t(\boldsymbol{\beta}) = \mathrm{P}(Y_t = 1|\mathcal{F}_{t-1}) = \mathrm{F}(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}),$$

- $\boldsymbol{\beta}$ je k-rozmerný vektor parametrov,
- $\mathcal{F}_{t-1} = \sigma(Y_{t-1}, Y_{t-2},..., \boldsymbol{Z}_{t-1}, \boldsymbol{Z}_{t-2},...)$ predstavuje informácie známe v čase $t-1$.

Zrejme $\mathcal{F}_{t-1} \subset \mathcal{F}_t$. Predpokladáme, že F je diferencovateľná distribučná funkcia s hustotou $f = F'$. Z binárnej povahy veličín Y_t vyplýva

$$P_\beta(Y_t = y_t|\mathcal{F}_{t-1}) = [\pi_t(\boldsymbol{\beta})]^{y_t}[1 - \pi_t(\boldsymbol{\beta})]^{1-y_t}, y_t \in \{0,1\}.$$

Z (1.10) dostávame

$$\begin{aligned} PL(\boldsymbol{\beta}) &= PL(\boldsymbol{\beta}; y_1,...,y_N) = \\ &= \prod_{t=1}^{N} [\pi_t(\boldsymbol{\beta})]^{y_t} [1 - \pi_t(\boldsymbol{\beta})]^{1-y_t} = \\ &= \prod_{t=1}^{N} [\mathrm{F}(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})]^{y_t}[1 - \mathrm{F}(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})]^{1-y_t}. \end{aligned} \tag{1.11}$$

Pre maximalizáciu vierohodnostnej funkcie používame logaritmickú vierohodnosť

$$l(\boldsymbol{\beta}) = \ln PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{y_t \ln [F(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})] + (1 - y_t) \ln [1 - F(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})]\}. \tag{1.12}$$

1.3 Logistická autoregresia

Autoregresný model vychádza z predpokladu, že každá hodnota v časovom rade závisí na predchádzajúcich hodnotách toho radu. AR(p), autoregresný model rádu p, definujeme podľa knihy Kedem a Fokianos (2002) predpisom

$$X_t = \gamma_0 + \gamma_1 X_{t-1} + ... + \gamma_p X_{t-p} + \lambda\varepsilon_t, \tag{1.13}$$

$\lambda > 0$ je konštanta, ε_t sú nezávislé rovnako rozdelené náhodné veličiny so štandardizovaným logistickým rozdelením s hustotou (1.6). Zvolíme prahovú hodnotu $r \in (-\infty,\infty)$ a definujeme binárny časový rad

$$Y_t = \mathbb{I}_{[X_t \geq r]}. \tag{1.14}$$

Ak označíme

$$\begin{aligned} \boldsymbol{Z}_{t-1} &= (1,X_{t-1},...,X_{t-p})', \\ \boldsymbol{\beta} &= \frac{1}{\lambda}(\gamma_0 - r,\gamma_1,...,\gamma_p)', \end{aligned}$$

máme

$$\boldsymbol{\beta}'\boldsymbol{Z}_{t-1} = \frac{1}{\lambda}(\gamma_0 - r + \gamma_1 X_{t-1} + ... + \gamma_p X_{t-p}).$$

S využitím (1.5) ďalej dostávame

$$\begin{aligned}\pi_t(\boldsymbol{\beta}) &= \mathrm{P}(Y_t = 1|\mathcal{F}_{t-1}) = \mathrm{P}(X_t \geq r|X_{t-1},...,X_{t-p}) = \\ &= \mathrm{P}(\gamma_0 + \gamma_1 X_{t-1} + ... + \gamma_p X_{t-p} + \lambda\varepsilon_t \geq r|X_{t-1},...,X_{t-p}) = \\ &= \mathrm{P}(\varepsilon_t \geq \frac{r - \gamma_0 - \gamma_1 X_{t-1} - ... - \gamma_p X_{t-p}}{\lambda}|X_{t-1},...,X_{t-p}).\end{aligned}$$

ε_t má distribučnú funkciu

$$\begin{aligned}\mathrm{F}(x) &= \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}, \\ 1 - \mathrm{F}(x) &= 1 - \frac{e^x}{1+e^x} = \frac{1}{1+e^x}.\end{aligned}$$

To znamená

$$\begin{aligned}\mathrm{P}(Y_t = 1|\mathcal{F}_{t-1}) &= 1 - \mathrm{F}\left(\frac{1}{\lambda}\left(r - \gamma_0 - \gamma_1 X_{t-1} - ... - \gamma_p X_{t-p}\right)\right) = \\ &= \frac{1}{1 + \exp\left[\frac{-1}{\lambda}\left(\gamma_0 - r + \gamma_1 X_{t-1} + ... + \gamma_p X_{t-p}\right)\right]} = \\ &= \mathrm{F}(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}) = \frac{1}{1+e^{-\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}}}.\end{aligned} \tag{1.15}$$

1.4 Odhad parametrov

Aby sme našli odhad vektora parametrov $\boldsymbol{\beta}$, odvodili sme si logaritmickú vierohodnostnú funkciu (1.12), do ktorej teraz dosadíme z (1.15).

$$l(\boldsymbol{\beta}) = \ln PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} y_t \ln \frac{1}{1+e^{-\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}}} + \sum_{t=1}^{N}(1-y_t)\ln\left(1 - \frac{1}{1+e^{-\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}}}\right). \tag{1.16}$$

Položíme

$$\nabla \ln PL(\boldsymbol{\beta}) = 0, \tag{1.17}$$

čím získame sústavu vierohodnostných rovníc. Konkrétnu podobu tejto sústavy odvodíme pre rád autoregresie $p = 1$. Pre model (1.13) to znamená

$$\begin{aligned}X_t &= \gamma_0 + \gamma_1 X_{t-1} + \lambda\varepsilon_t, \\ \boldsymbol{\beta}'\boldsymbol{Z}_{t-1} &= \frac{1}{\lambda}(\gamma_0 - r + \gamma_1 X_{t-1}) = \beta_0 + \beta_1 X_{t-1}, \\ \beta_0 &= \frac{\gamma_0 - r}{\lambda}, \\ \beta_1 &= \frac{\gamma_1}{\lambda}.\end{aligned} \tag{1.18}$$

Budeme derivovať logaritmickú vierohodnostnú funkciu podľa zložiek vektoru $\boldsymbol{\beta}$ a derivácie položíme rovné nule. Pri konkrétnych hodnotách y_t, x_{t-1} je

$$l(\boldsymbol{\beta}) = \sum_{t=1}^{N} y_t \log \frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}} + \sum_{t=1}^{N}(1-y_t)\log\left(1-\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}}\right).$$

$$\begin{aligned} 0 = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{t=1}^{N} y_t \cdot \left(1+e^{-\beta_0-\beta_1 x_{t-1}}\right)\cdot(-1)\cdot\frac{e^{-\beta_0-\beta_1 x_{t-1}}}{\left(1+e^{-\beta_0-\beta_1 x_{t-1}}\right)^2}\cdot(-1)+ \\ &+ \sum_{t=1}^{N}(1-y_t)\cdot\frac{1+e^{-\beta_0-\beta_1 x_{t-1}}}{e^{-\beta_0-\beta_1 x_{t-1}}}\cdot\frac{e^{-\beta_0-\beta_1 x_{t-1}}}{(1+e^{-\beta_0-\beta_1 x_{t-1}})^2}\cdot(-1) = \\ &= \sum_{t=1}^{N} y_t\left(1-\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}}\right)+\sum_{t=1}^{N}(y_t-1)\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}} = \\ &= \sum_{t=1}^{N}\left(y_t-\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}}\right). \end{aligned} \tag{1.19}$$

$$\begin{aligned} 0 = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} &= \sum_{t=1}^{N} y_t \cdot \left(1+e^{-\beta_0-\beta_1 x_{t-1}}\right)\cdot(-1)\cdot\frac{e^{-\beta_0-\beta_1 x_{t-1}}}{\left(1+e^{-\beta_0-\beta_1 x_{t-1}}\right)^2}\cdot(-1)\cdot x_{t-1}+ \\ &+ \sum_{t=1}^{N}(1-y_t)\cdot\frac{1+e^{-\beta_0-\beta_1 x_{t-1}}}{e^{-\beta_0-\beta_1 x_{t-1}}}\cdot\frac{e^{-\beta_0-\beta_1 x_{t-1}}}{\left(1+e^{-\beta_0-\beta_1 x_{t-1}}\right)^2}\cdot(-1)\cdot x_{t-1} = \\ &= \sum_{t=1}^{N} y_t\cdot x_{t-1}\left(1-\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}}\right)+\sum_{t=1}^{N}(y_t-1)\,x_{t-1}\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}} \\ &= \sum_{t=1}^{N} x_{t-1}\left(y_t-\frac{1}{1+e^{-\beta_0-\beta_1 x_{t-1}}}\right). \end{aligned} \tag{1.20}$$

Rovnice (1.19) a (1.20) riešime s použitím vhodného softwareu vzhľadom k neznámym β_0 a β_1. Riešenia $\hat{\beta}_0$ a $\hat{\beta}_1$ sú hľadané odhady.

Asymptotické vlastnosti maximálne vierohodného odhadu sú zhrnuté v nasledujúcej vete, ktorej dôkaz môžeme nájsť v článku Kedem a Fokianos (1998).

Veta 1 (Kedem a Fokianos, 2002, str. 59)**.** *Odhad $\hat{\boldsymbol{\beta}}$ získaný metódou maximálnej vierohodnosti je skoro iste určený jednoznačne pre všetky dostatočne veľké N a pre $N\to\infty$ platí*

1.

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta},$$

2.

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \xrightarrow{\mathcal{D}} \mathcal{N}_k(\mathbf{0},\boldsymbol{G}^{-1}(\boldsymbol{\beta})),$$

3.

$$\sqrt{N}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) - \frac{1}{\sqrt{N}}\boldsymbol{G}^{-1}(\boldsymbol{\beta})\boldsymbol{S}_N(\boldsymbol{\beta}) \xrightarrow{P} \mathbf{0}.$$

Vysvetlíme symboly použité vo vete 1. V bode 3 je $\boldsymbol{S}_N(\boldsymbol{\beta}) \equiv \nabla \ln PL(\boldsymbol{\beta})$. V knihe Kedem a Fokianos (2002, str. 60) je derivácia uvedená v zjednodušenej vektorovej forme, ktorej platnosť ukážeme pre prípad $p = 1$:

$$\nabla \ln PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \left[Y_t - \pi_t(\boldsymbol{\beta}) \right] = \sum_{t=1}^{N} \begin{bmatrix} 1 \\ X_{t-1} \end{bmatrix} \left(Y_t - \frac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right) =$$
$$= \sum_{t=1}^{N} \begin{bmatrix} Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \\ X_{t-1} \left(Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right) \end{bmatrix}. \tag{1.21}$$

Tento tvar zodpovedá deriváciám podľa zložiek vektoru $\boldsymbol{\beta}$ (1.19), kde sme použili pozorované hodnoty y_t, x_{t-1} náhodných veličín Y_t, X_{t-1}.

Označíme $\boldsymbol{G}_N(\boldsymbol{\beta}) \equiv \nabla\nabla'(-\ln PL(\boldsymbol{\beta}))$, čo je pre $p = 1$ matica s rozmermi 2x2. Odvodíme si jej prvky

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial {\beta_0}^2} = -\frac{\partial \left(\sum_{t=1}^{N} Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right)}{\partial \beta_0} = \sum_{t=1}^{N} \frac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2},$$
$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1 \beta_0} = -\frac{\partial \left(\sum_{t=1}^{N} X_{t-1} \left(Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right) \right)}{\partial \beta_0} = \sum_{t=1}^{N} X_{t-1} \frac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2},$$
$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0 \beta_1} = -\frac{\partial \left(\sum_{t=1}^{N} Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right)}{\partial \beta_1} = \sum_{t=1}^{N} X_{t-1} \frac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2},$$
$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial {\beta_1}^2} = -\frac{\partial \left(\sum_{t=1}^{N} X_{t-1} \left(Y_t - \dfrac{1}{1 + e^{-\beta_0 - \beta_1 X_{t-1}}} \right) \right)}{\partial \beta_1} = \sum_{t=1}^{N} X_{t-1}^2 \frac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2}.$$

Vektorový tvar možno opäť nájsť v Kedem a Fokianos (2002, str. 60). Ukážeme, že pre $p = 1$ má matica $\boldsymbol{G}_N(\boldsymbol{\beta})$ vyššie uvedené prvky.

$$\boldsymbol{G}_N(\boldsymbol{\beta}) = \nabla\nabla'(-\ln PL(\boldsymbol{\beta})) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \boldsymbol{Z}'_{t-1} \pi_t(\boldsymbol{\beta}) \left[1 - \pi_t(\boldsymbol{\beta}) \right] =$$
$$= \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \boldsymbol{Z}'_{t-1} \frac{e^{\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}}}{(1 + e^{\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}})^2} =$$
$$= \sum_{t=1}^{N} \begin{bmatrix} 1 \\ X_{t-1} \end{bmatrix} \begin{bmatrix} 1 & X_{t-1} \end{bmatrix} \frac{e^{\beta_0 + \beta_1 X_{t-1}}}{(1 + e^{\beta_0 + \beta_1 X_{t-1}})^2} =$$
$$= \sum_{t=1}^{N} \begin{bmatrix} 1 & X_{t-1} \\ X_{t-1} & X_{t-1}^2 \end{bmatrix} \frac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2} =$$
$$= \sum_{t=1}^{N} \begin{bmatrix} \dfrac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2} & X_{t-1} \dfrac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2} \\ X_{t-1} \dfrac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2} & X_{t-1}^2 \dfrac{e^{-\beta_0 - \beta_1 X_{t-1}}}{(1 + e^{-\beta_0 - \beta_1 X_{t-1}})^2} \end{bmatrix}. \tag{1.22}$$

Pre $\boldsymbol{G}(\boldsymbol{\beta})$ z 2. a 3. bodu Vety 1 platí

$$\frac{\boldsymbol{G}_N(\boldsymbol{\beta})}{N} \xrightarrow[N\to\infty]{\mathrm{P}} \boldsymbol{G}(\boldsymbol{\beta}).$$

Podľa Kedem a Fokianos (2002, str. 65) je

$$\boldsymbol{G}(\boldsymbol{\beta}) = E\left[\frac{e^{\boldsymbol{\beta}'\boldsymbol{Z}}}{(1+e^{\boldsymbol{\beta}'\boldsymbol{Z}})^2}\cdot \boldsymbol{Z}\boldsymbol{Z}'\right] = E[f_l(\boldsymbol{\beta}'\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{Z}']. \tag{1.23}$$

Veta 1 má praktické využitie v konštrukcii intervalových odhadov pre $\pi_t(\boldsymbol{\beta})$ pomocou $\boldsymbol{Z}_{t-1}$. Použitím delta metódy (Rao, 1973) v druhom bode vety 1 dostávame

$$\begin{aligned}\sqrt{N}[\pi_t(\hat{\boldsymbol{\beta}}) - \pi_t(\boldsymbol{\beta})] &\xrightarrow[N\to\infty]{\mathcal{D}} \mathcal{N}(0, \gamma'\boldsymbol{G}^{-1}(\boldsymbol{\beta})\gamma), \\ \gamma = \nabla\pi_t(\boldsymbol{\beta}) = \nabla \mathrm{F}_l(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}) &= f_l(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})\boldsymbol{Z}_{t-1},\end{aligned} \tag{1.24}$$

F_l je distribučná funkcia a f_l hustota štandardizovaného logistického rozdelenia.

Náhodná veličina $\frac{\sqrt{N}}{\sqrt{\gamma'\boldsymbol{G}^{-1}(\boldsymbol{\beta})\gamma}}\left[\pi_t(\hat{\boldsymbol{\beta}}) - \pi_t(\boldsymbol{\beta})\right]$ má teda asymptoticky rozdelenie N(0,1), dosadením za γ máme

$$|\pi_t(\hat{\boldsymbol{\beta}}) - \pi_t(\boldsymbol{\beta})| < u_{1-\alpha/2}\frac{f_l(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})}{\sqrt{N}}\sqrt{\boldsymbol{Z}'_{t-1}\boldsymbol{G}^{-1}(\boldsymbol{\beta})\boldsymbol{Z}_{t-1}}$$

a dostávame hranice asymptotického 100(1-α)% intervalového odhadu $\pi_t(\boldsymbol{\beta})$

$$\pi_t(\hat{\boldsymbol{\beta}}) \pm u_{1-\alpha/2}\frac{f_l(\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_{t-1})}{\sqrt{N}}\cdot\sqrt{\boldsymbol{Z}'_{t-1}\boldsymbol{G}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{Z}_{t-1}}\,, \tag{1.25}$$

kde $u_{1-\alpha/2}$ označuje $1-\alpha/2$ kvantil štandardizovaného normálneho rozdelenia.

2. Simulačná štúdia

V praktickej časti sme realizovali štúdiu založenú na Monte Carlo simulácii, ktorej hlavným cieľom je ilustrácia teoretických výsledkov z kapitoly 1 a ich overenie. Pre počítačovú implementáciu sme zvolili voľne dostupný programovací jazyk R (R Core Team, 2019), ktorý je v súčasnosti najpoužívanejším programovacím jazykom v oblasti štatistiky a je mimoriadne vhodným pre štatistický výskum akýchkoľvek časových radov s využitím simulácii (McLeod a kol., 2012). Samotné naprogramovanie algoritmu simulácie sme realizovali vo voľne dostupnom vývojovom prostredí RStudio.

2.1 Algoritmus Monte Carlo simulácie

Naprogramovali sme algoritmus pre Monte Carlo simuláciu AR(1) procesu:

$$X_t = \gamma_1 X_{t-1} + \varepsilon_t, \tag{2.1}$$

kde $\varepsilon_t, t = 1,...,N$, sú chyby zo štandardizovaného logistického rozdelenia (1.5). V (1.18) volíme parameter $\lambda = 1$ a $\gamma_0 = 0$. Odtiaľ máme v (1.18) $\beta_0 = \frac{\gamma_0 - r}{\lambda} = -r$ a $\beta_1 = \frac{\gamma_1}{\lambda} = \gamma_1$.

Algoritmus možno prehľadne zapísať v nasledovnom tvare zloženom z troch častí:

- *Definovanie vstupných parametrov:* γ_1, N, B.
- *Definovanie pomocných funkcií pre simuláciu:* `generuj_ar1`, `generuj_Y`, `log_vierohodnost`.
- *Monte Carlo simulácia:* pre $k = 1, \ldots, B$ opakovanie nasledovných krokov
  1. Generovanie AR(1) realizácie časového radu dĺžky N a parametrom γ_1: $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)'$.
  2. Výpočet mediánu r zložiek $\boldsymbol{X}$.
  3. Generovanie binárneho procesu Y_t z $\boldsymbol{X}$ s prahom r.
  4. Výpočet odhadu β_1 metódou maximálnej vierohodnosti.

V prvej časti algoritmu sme teda nastavili vstupné parametre simulácie. Pre koeficient γ_1 sme zvolili tri hodnoty 0,1; 0,5 a 0,9. Podobne pre parameter N určujúci dĺžku radu sme tiež stanovili tri hodnoty 100, 500 a 1000. Všetky kombinácie parametrov tak vytvorili deväť skúmaných prípadov.

V ďalšej časti sme definovali pomocné funkcie na generovanie procesu X_t, generovanie príslušných hodnôt procesu Y_t definovaného vzťahom (1.14), kde sme ako prahovú hodnotu r volili medián vygenerovaného radu $\{X_t; t = 1,...,N\}$ a log-vierohodnostnú funkciu podľa (1.16). Pri generovaní logistického rozdelenia sme využili v R existujúcu funkciu `rlogis`.

Hlavné telo algoritmu tvorí simulácia pre jednotlivé hodnoty $\gamma_1 = \beta_1$ a dĺžky radu N. V nej sme generovali $B = 1000$ realizácii časového radu pre všetky skúmané prípady. Z časového hľadiska výpočet výsledkov na základe Monte Carlo simulácie pre všetky kombinácie parametrov trval na bežnom počítači niekoľko minút.

Výstupom je zhrnutie výsledkov vo forme tabuľky a grafov. Kompletný algoritmus je na CD vo formáte `rmd` a `pdf`.

2.2 Výsledky simulácie

Všetky výsledky simulácie sú zhrnuté vo forme tabuľky 2.1 reprezentujúcej prehľadnú číselnú sumarizáciu a v zodpovedajúcich obrázkoch 2.1, 2.2 a 2.3 predstavujúcich grafickú sumarizáciu.

Dĺžka radu	Priemer odhadov $\hat{\beta}_1$			Smerodajná odchýlka odhadov $\hat{\beta}_1$		
	$\beta_1 = 0{,}1$	$\beta_1 = 0{,}5$	$\beta_1 = 0{,}9$	$\beta_1 = 0{,}1$	$\beta_1 = 0{,}5$	$\beta_1 = 0{,}9$
100	0,1018	0,5205	0,9386	0,1201	0,1349	0,1468
500	0,1001	0,5064	0,9173	0,0486	0,0587	0,0787
1000	0,0987	0,5026	0,9128	0,0354	0,0416	0,0553

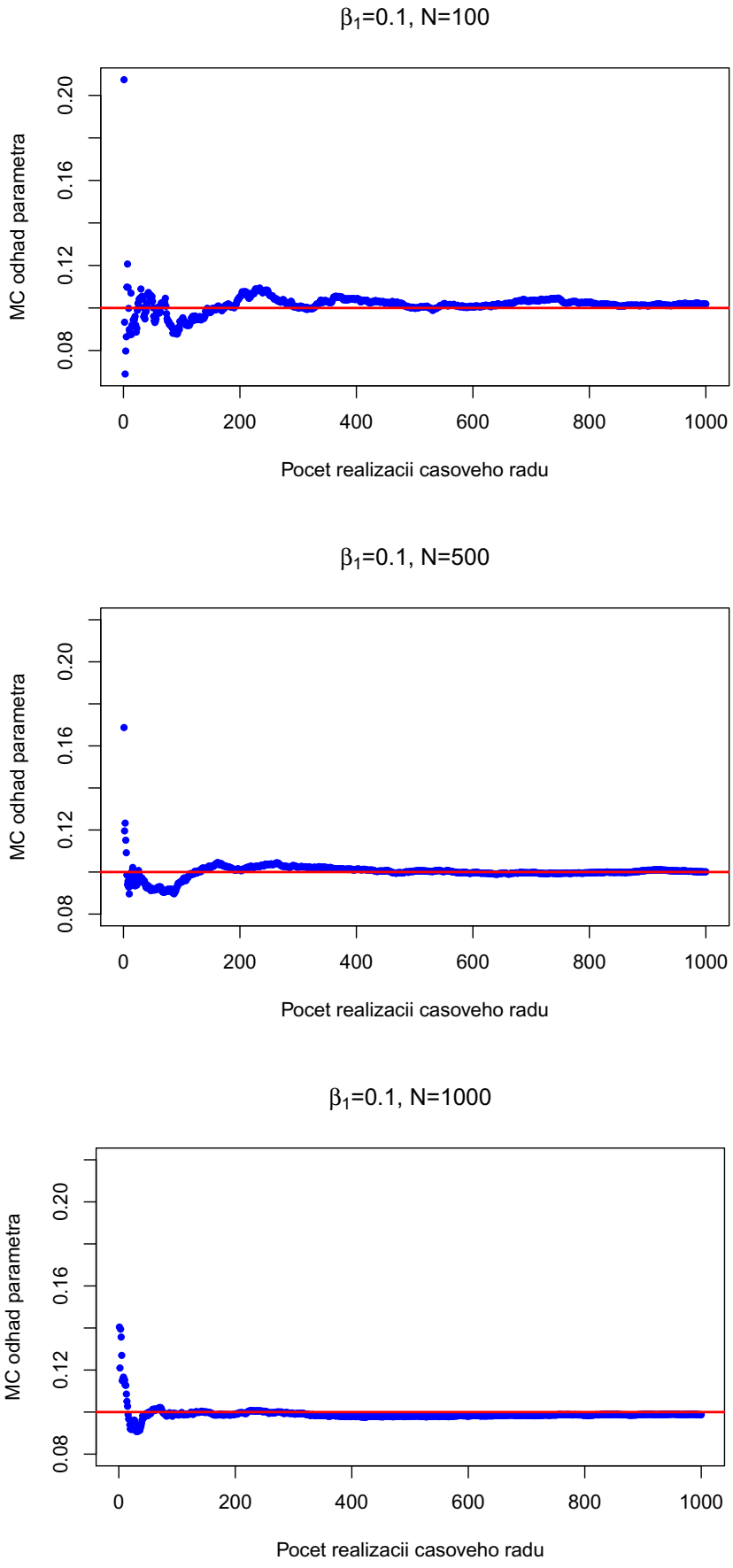
Tabuľka 2.1: Výsledky simulácie

V tabuľke 2.1 sme pre každú zvolenú hodnotu dĺžky radu $N = 100, 500, 1000$ a každú z troch zvolených skutočných hodnôt parametra $\beta_1 = 0{,}1; 0{,}5, 0{,}9$ na základe $B = 1000$ opakovaní spočítali základné charakteristiky odhadu $\hat{\beta}_1$: aritmetický priemer a smerodajnú odchýlku.

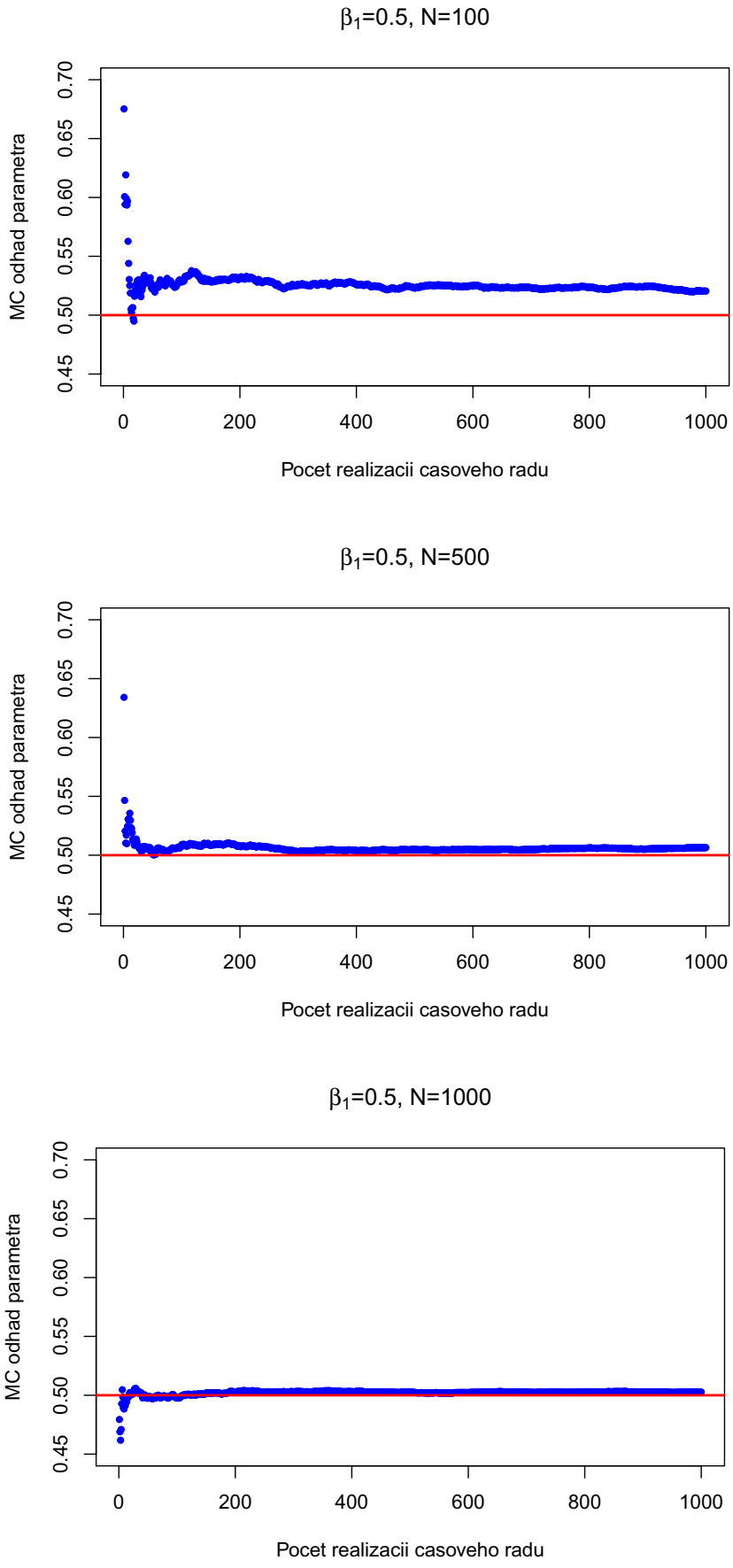
Všeobecne aritmetický priemer z k opakovaní simulácie, $k = 1, \ldots, B$, budeme kvôli jednoduchosti vyjadrovania nazývať *Monte Carlo odhad* parametra β_1, resp. v skratke *MC odhad* parametra β_1. Obdobne pre smerodajnu odchýlku z k opakovaní zavedieme pojem *Monte Carlo smerodajná odchýlka* odhadu parametra β_1 alebo *MC smerodajná odchýlka* parametra.

Výsledky ukazujú, že MC odhady parametra $\beta_1 = \gamma_1$ metódou maximálnej vierohodnosti sú blízke skutočným hodnotám pre všetky tri sledované dĺžky radu. MC odhady sa pohybujú okolo skutočných hodnôt, pričom s rastúcim N majú tendenciu byť bližšie ku skutočnej hodnote parametra. Táto skutočnosť naznačuje dostatočne rýchlu konvergenciu odhadu $\hat{\beta}_1$ ku skutočnej hodnote β_1 s rastúcim N.

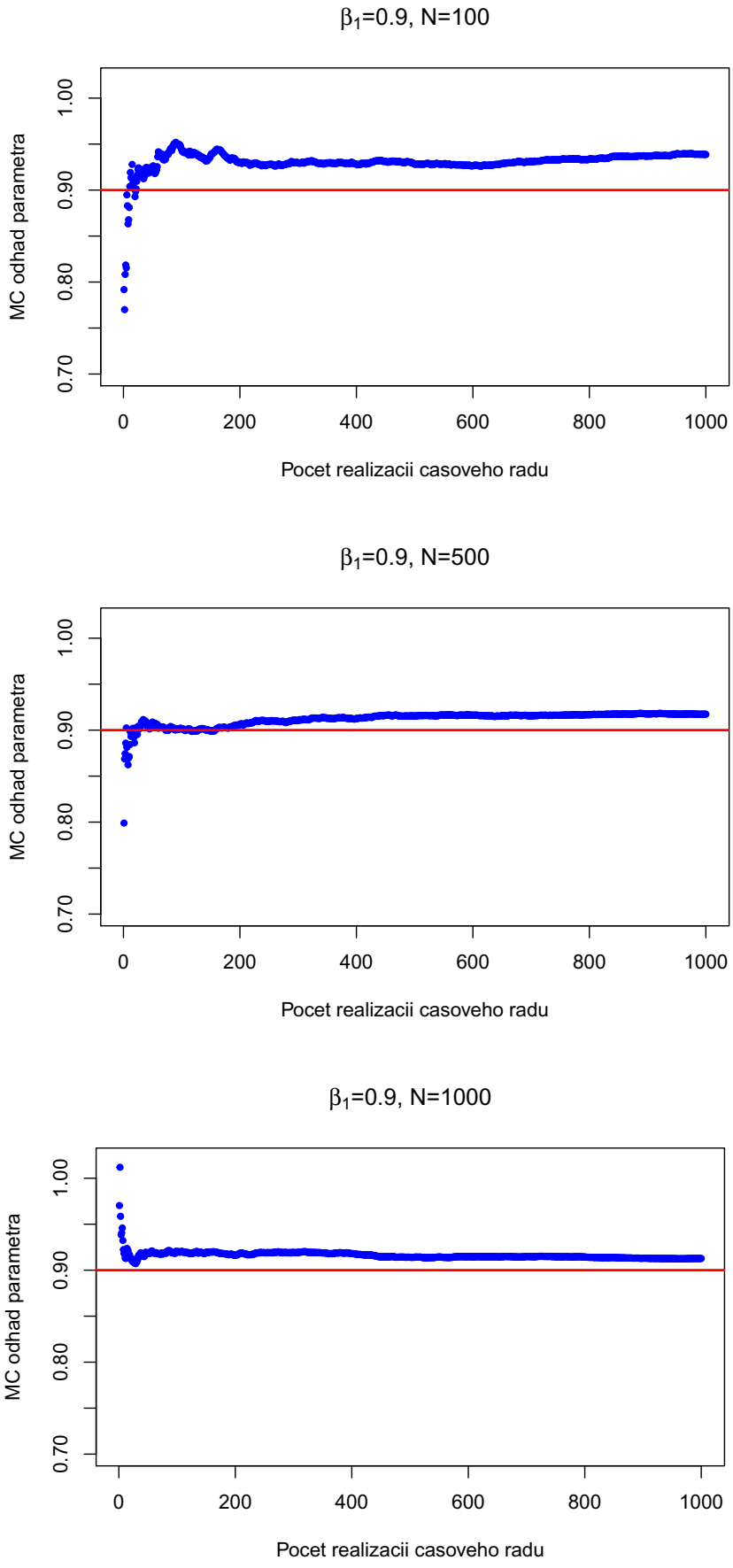
S rastúcou dĺžkou radu N klesajú aj smerodajné odchýlky, čo potvrdzuje konzistentnosť odhadov. Výsledky simulácie sú teda v súlade s našimi teoretickými poznatkami z vety 1.



Obrázok 2.1: Graf závislosti MC odhadu parametra na počte opakovaní k v simulácii pre $\beta_1 = 0{,}1$



Obrázok 2.2: Graf závislosti MC odhadu parametra na počte opakovaní k v simulácii pre $\beta_1 = 0{,}5$



Obrázok 2.3: Graf závislosti MC odhadu parametra na počte opakovaní k v simulácii pre $\beta_1 = 0{,}9$

Grafy na obrázkoch 2.1, 2.2 a 2.3 zobrazujú, ako sa vyvíjala hodnota MC odhadu β_1 s rastúcim počtom opakovaní $k = 1,...,B$ Monte Carlo simulácie. Inými slovami dané grafy znázorňujú závislosť MC odhadu parametra β_1 na počte generovaných realizácií k nášho AR(1) procesu. Súčasne nám to dovoľuje vidieť aj rýchlosť konvergencie MC odhadu ku skutočnej hodnote parametra β_1 s rastúcim počtom simulácií k, t.j. po koľkých opakovaniach k sa zhruba hodnota MC odhadu stabilizuje.

Pri všetkých hodnotách parametov $\beta_1 = 0{,}1; 0{,}5; 0{,}9$ vidíme, že s rastúcou dĺžkou N sa MC odhad stabilizuje rýchlejšie, t.j. pri menšom počte opakovaní k. To znamená, že najrýchlejšie je to pri dĺžke radu $N = 1000$ pre všetky hodnoty prarametra a to približne pre $k = 100$.

3. Aplikácia na reálne dáta

V praxi sa s analýzou a predikciou budúcich hodnôt binárnych časových radov môžeme stretnúť v rôznych oblastiach. Napríklad z realizácie časového radu denného úhrnu zrážok možno získať realizáciu binárneho časového radu s hodnotami 1 - prší, 0 - neprší. Zo záznamu trvania erupcie gejzíru dostaneme binárny časový rad s hodnotami 1 - trvanie aspoň 3 minúty, 0 - krátke trvanie pod 3 minúty (Kedem a Fokianos, 2002).

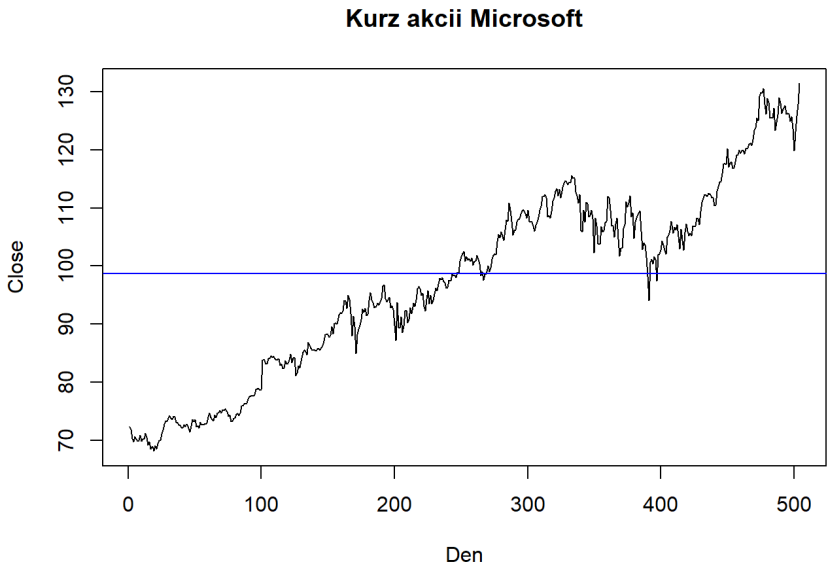
Na analýzu reálneho časového radu nula-jednotkovej povahy sme sa rozhodli vyhľadať dáta vyjadrujúce cenu akcií nejakej spoločnosti. Sledovať, či hodnoty ceny akcií prekročia, resp. neprekročia istú zvolenú hodnotu môže mať význam pri rozhodovaní sa, či budeme akcie v nasledujúcom období predávať alebo kupovať.

Za východzie hodnoty sme zvolili historické dáta kurzu akcií spoločnosti Microsoft v čase od 07.06.2017 do 07.06.2019. Údaje sme získali z Nasdaq (2019). Za toto obdobie máme N=504 pozorovaní z dní, kedy sa na burze obchodovalo.

Z realizácie $(x_0, x_1,...,x_{503})'$ časového radu cien akcií $\{X_t\}$ sme určili medián $r = 98{,}645$ ako prahovú hodnotu pre binárny časový rad $\{Y_t\}$. Vektor realizácií binárneho časového radu sme označili ako $\boldsymbol{Y_1}$, pričom jeho zložky sme získali podľa vzťahu

$$Y_{1t} = \begin{cases} 1 & \text{pre } X_t \geq r \\ 0 & \text{pre } X_t < r, \end{cases}$$

Analyzované dáta s vypočítaným mediánom sú znázornené na obrázku 3.1.



Obrázok 3.1: Graf kurzu akcií firmy Microsoft v období 07.06.2017 – 07.06.2019 s vyznačeným mediánom

Kvôli overeniu teoretických poznatkov z prvej kapitoly na reálnych dátach, sme na vytvorenie binárneho časového radu $\{Y_t\}$ použili odhad $\pi_t(\hat{\boldsymbol{\beta}})$ pravdepodobnosti $\pi_t(\boldsymbol{\beta}) = \mathrm{P}(X_t \geq r|X_{t-1})$, odvodenej v (1.15). Pre odlíšenie od realizácie

$\boldsymbol{Y_1}$, ktorú sme získali priamo z nameraných hodnôt ceny akcií, sme označili vektor realizácií ako $\boldsymbol{Y_2}$. Bodové odhady sme pritom využili nasledovne:

$$\hat{Y}_{2t} = \begin{cases} 1 & \text{pre } \pi_t(\hat{\boldsymbol{\beta}}) \geq 0{,}5 \\ 0 & \text{pre } \pi_t(\hat{\boldsymbol{\beta}}) < 0{,}5. \end{cases}$$

Zložky parametra $\hat{\boldsymbol{\beta}}$ sme odhadli metódou maximálnej vierohodnosti z realizácie časového radu $\{X_t\}_{t=0}^{N-1}$. Dostali sme hodnoty $\hat{\boldsymbol{\beta}} = (-96{,}69; 0{,}98)'$. Následne sme urobili predikciu budúcej hodnoty binárneho časového radu Y_{504}, ktorá podľa očakávania vyšla rovná jednej.

Asymptotický 95% intervalový odhad pre $\pi_t(\boldsymbol{\beta})$, ktorý sme odvodili v prvej kapitole (vzťah (1.25))

$$\pi_t(\hat{\boldsymbol{\beta}}) \pm u_{0{,}975} \frac{f_l(\hat{\boldsymbol{\beta}}' \boldsymbol{Z}_{t-1})}{\sqrt{N}} \cdot \sqrt{\boldsymbol{Z}'_{t-1} \boldsymbol{G}^{-1}(\hat{\boldsymbol{\beta}}) \boldsymbol{Z}_{t-1}}\,, \tag{3.1}$$

sme využili ako tretí spôsob získania hodnôt binárneho časového radu, resp. jeho odhadu. Vektor realizácie časového radu sme v súlade s predchádzajúcim značením označili ako $\boldsymbol{Y_3}$. Ak označíme dolnú hranicu intervalu D a hornú hranicu H. Potom zložky vektora $\boldsymbol{Y_3}$ sa riadia kritériom

$$\hat{Y}_{3t} = \begin{cases} 1 & \text{pre } 0.5 < \mathrm{D} < \mathrm{H} < 1 \\ 0 & \text{pre } 0 < \mathrm{D} < \mathrm{H} < 0{,}5 \\ - & \text{pre } 0 < \mathrm{D} < 0{,}5 < \mathrm{H} < 1. \end{cases}$$

Zložky vektora $\hat{\boldsymbol{\beta}}$ sme odhadli rovnakým spôsobom ako pri bodovom odhade. $\boldsymbol{G}^{-1}(\hat{\boldsymbol{\beta}})$ sme odhadli ako

$$\boldsymbol{G}^{-1}(\hat{\boldsymbol{\beta}}) \approx \frac{1}{N} \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \boldsymbol{Z}'_{t-1} \pi_t(\hat{\boldsymbol{\beta}})(1 - \pi_t(\hat{\boldsymbol{\beta}})),$$

$$\boldsymbol{Z}_{t-1} = (1, X_{t-1})'.$$

Opäť sme našli aj predikciu nasledujúcej hodnoty radu Y_{504}. Predikovaná hodnota 1 sa zhodovala s predikciou na základe bodových odhadov.

Nasledujúca tabuľka 3.1 zhŕňa, koľkokrát sa odhady hodnôt binárneho časového radu získané pomocou bodových odhadov ($\boldsymbol{Y_2}$), resp. na základe intervalových odhadov ($\boldsymbol{Y_3}$) zhodli so skutočnými pozorovaniami z $\boldsymbol{Y_1}$. V poslednom stĺpci je uvedená pre oba spôsoby predikcia pre nasledujúci časový okamih. Vidíme, že percentuálna zhoda bola vysoká v oboch prípadoch a teda použitý model je vyhovujúci. Môžeme teda očakávať, že v nasledujúcom časovom okamihu bude hodnota ceny akcie firmy Microsoft nad mediánom.

	Zhoda-počet	Zhoda-percento	Predikcia
$\boldsymbol{Y_2}$	493	98,01	1
$\boldsymbol{Y_3}$	450	89,46	1

Tabuľka 3.1: Porovnanie skutočných hodnôt binárneho časového radu s odhadmi pre r = medián = 98,645

Z praktického hľadiska by mohlo mať zmysel voliť aj inú prahovú hodnotu ako medián. Napríklad by sme mohli sledovať, kedy bude cena akcií dostatočne vysoká, aby sa oplatilo akcie predať. Vyskúšali sme viacero hodnôt. Tabuľky 3.2, 3.3 a 3.4 sumarizujú výsledky pre prahové hodnoty $r = 120$, $r = 125$ a $r = 128$. Postupne sme dostali odhady $\hat{\boldsymbol{\beta}}_{120} = (-105{,}45; 0{,}88)'$, $\hat{\boldsymbol{\beta}}_{125} = (-115{,}85; 0{,}93)'$, $\hat{\boldsymbol{\beta}}_{128} = (-85{,}10; 0{,}66)'$.

	Zhoda-počet	Zhoda-percento	Predikcia
$\boldsymbol{Y_2}$	498	99,01	1
$\boldsymbol{Y_3}$	475	94,43	1

Tabuľka 3.2: Porovnanie skutočných hodnôt binárneho časového radu s odhadmi pre $r = 120$

	Zhoda-počet	Zhoda-percento	Predikcia
$\boldsymbol{Y_2}$	498	99,01	1
$\boldsymbol{Y_3}$	477	94,83	1

Tabuľka 3.3: Porovnanie skutočných hodnôt binárneho časového radu s odhadmi pre $r = 125$

	Zhoda-počet	Zhoda-percento	Predikcia
$\boldsymbol{Y_2}$	498	99,01	1
$\boldsymbol{Y_3}$	494	98,21	0

Tabuľka 3.4: Porovnanie skutočných hodnôt binárneho časového radu s odhadmi pre $r = 128$

Zhoda odhadov hodnôt binárneho časového radu $\{Y_t\}$, či cena akcie firmy Microsoft bude nad istú prahovú hodnotu r alebo nebude, so skutočnými pozorovaniami, bola vysoká pre všetky nami zvolené prahové hodnoty.

Skutočná hodnota akcie firmy Microsoft dňa 10.06.2019 bola 132,6, čo znamená, že predikcie nám vyšli v každom prípade okrem posledného intervalového odhadu pre $r = 128$.

Podobne ako Monte Carlo simuláciu v druhej kapitole, aj celé spracovanie reálnych dát analyzovaných v tretej kapitole, sme previedli pomocou softvéru R (R Core Team, 2019). Okrem nami definovaných pomocných funkcií `generuj_Y` a `log_vierohodnost`, použitých v simulácii, sme definovali ďalšie dve funkcie `pravdepodobnost` a `IS_beta`. Pomocou nich sme dokázali vygenerovať hodnoty binárneho časového radu na základe reálnych dát a tiež nájsť bodové a intervalové odhady potrebné k odhadovaniu hodnôt spomínaného binárneho časového radu. K odhadu parametra $\boldsymbol{\beta}$ bolo potrebné riešiť sústavu vierohodnostných rovníc (1.19) a (1.20). Využili sme zabudovanú optimalizačnú funkciu `nlm`, ktorá umožňuje nájsť minimum nelineárnej funkcie, pričom logaritmická vierohodnosť

u nás figuruje s opačným znamienkom. Na overenie výsledkov sme použili funkcie `nleqslv` z balíka nleqslv (Hasselman, 2018) a `fsolve` z balíka pracma (Borchers, 2019), ktoré umožňujú riešiť sústavu nelineárnych rovníc.

Dáta vo formáte `xlsx` a kompletný kód s komentármi je vo formáte `rmd`, `pdf` a `html` elektronickou prílohou tejto práce.

Záver

V tejto bakalárskej práci sme sa venovali modelovaniu binárnych časových radov. V súlade s vytýčenými cieľmi sme prácu rozdelili na 3 kapitoly.

V prvej kapitole sme sa venovali logistickému rozdeleniu, ktoré má praktické využitie práve pri modelovaní binárnych časových radov. Okrem skúmania hustoty a distribučnej funkcie sme odvodili základné charakteristiky polohy a variability rozdelenia – strednú hodnotu a rozptyl. V druhej podkapitole sme sa zaoberali metódou maximálnej vierohodnosti, konkrétne parciálnou vierohodnosťou, ktorá umožňuje pracovať so závislými dátami, čo je práve prípad časových radov. V ďalšej časti sme predstavili autoregresný model s chybami zo štandardizovaného logistického rozdelenia, odvodili sme bodové a intervalové odhady parametrov pre model rádu jedna.

V druhej kapitole sme využili naštudované poznatky pri realizácii simulačnej štúdie. Naprogramovali sme algoritmus pre Monte Carlo simuláciu AR(1) procesu spomínaného vyššie, z ktorého sme odvodili binárny časový rad. Následne sme aplikovali metódu výpočtu odhadu parametra, s ktorou sme sa oboznámili v prvej kapitole. Pracovali sme s rôznymi dĺžkami simulovaného radu a sledovali sme zhodu odhadovaného parametra β_1 so skutočnou hodnotou pre tri rôzne hodnoty parametra. Výsledky simulácie ukázali, že MC odhady parametra metódou maximálnej vierohodnosti sú blízke skutočným hodnotám, pričom s rastúcou dĺžkou radu odhady konvergujú rýchlejšie k skutočnej hodnote a smerodajná odchýlka klesá. Potvrdili sa teda vlastnosti nestrannosť a konzistentnosť odhadov.

V poslednej kapitole sme uskutočnili analýzu reálnych dát, časového radu hodnôt akcií firmy Microsoft zaznamenávaných denne v rozpätí dvoch rokov. Z nameraných hodnôt sme získali niekoľko binárnych časových radov v závislosti na zvolenej podmienke. Hodnoty binárnych radov sme odhadovali aj na základe bodových a intervalových odhadov a porovnávali sme zhodu s pozorovanými dátami. Najnižšiu zhodu medzi hodnotami binárneho radu získaného z nameraných dát a hodnotami odhadovaného radu sme dosiahli pre prahovú hodnotu rovnú mediánu, konkrétne 98,01% pri využití bodových odhadov a 89,46% pri využití intervalových odhadov. Najvyššia zhoda 99,01% bola dosiahnutá pre prahovú hodnotu 128 v prípade bodových odhadov a 98,21% pri využití intervalových odhadov. Ďalej sme sa zaoberali predikciou. Konkrétne sme predikovali, či cena akcie v nasledujúcom obchodnom dni bude, resp. nebude vyššia ako zvolená prahová hodnota. Uvažovali sme štyri prahové hodnoty a bodové aj intervalové odhady. Predikcia sa nezhodovala so skutočnou hodnotou len v jednom prípade z ôsmich uvažovaných, a to pri prahovej hodnote 128 pri využití intervalových odhadov.

Celú praktickú časť práce sme previedli pomocou softvéru R. Využili sme nami vytvorené aj existujúce pomocné funkcie. Simulačná štúdia aj aplikácia na reálne dáta boli spracované vo forme dynamických dokumentov vo formáte `rmd`. Spolu s `html` verziou sú súčasťou elektronickej prílohy práce. Použité dáta nájde čitateľ tiež v prílohe vo formáte `xlsx`.

Zoznam použitej literatúry

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions*. URL `https://CRAN.R-project.org/package=pracma`. R package version 2.2.5.

Cipra, T. (2008). *Finanční ekonometrie*. Ekopress, Praha. ISBN 978-80-86929-43-9.

Hasselman, B. (2018). *nleqslv: Solve Systems of Nonlinear Equations*. URL `https://CRAN.R-project.org/package=nleqslv`. R package version 3.3.2.

Kedem, B. a Fokianos, K. (1998). Prediction and classification of non-stationary categorical time series. *Journal of Multivariate Analysis*, **67**(1), 277–296.

Kedem, B. a Fokianos, K. (2002). *Regression Models for Time Series Analysis*. 1. Wiley, New Yersey. ISBN 0-471-36355-3.

McLeod, A. I., Yu, H. a Mahdi, E. (2012). Time Series with R. In Rao, T. S., Rao, S. S. a Rao, C. R., editors, *Time Series Analysis: Methods and Applications*, volume 30, pages 661–712. Elsevier, Amsterdam. ISBN 978-0-444-53858-1.

Nasdaq (2019). Microsoft corporation common stock historical stock prices. `https://www.nasdaq.com/symbol/msft/historical`. [Online; accessed 2019-06-07].

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`.

Rao, C. (1973). *Linear Statistical Inference and Its Applications*. 2nd edition. Wiley, New York. ISBN 978-0471218753.

Zoznam obrázkov

Zoznam tabuliek