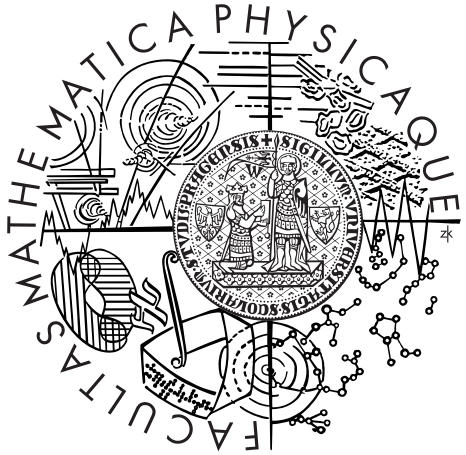
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Zuzana Dortová

Odhady parametrů založené na zaokrouhlených datech

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	prof. RNDr. Jiří Anděl, DrSc.
Studijní program:	Matematika
Studijní obor:	PMSE

Praha 2016

Děkuji prof. RNDr. Jiřímu Andělovi, DrSc. za trpělivé vedení práce a pomoc s jejím zpracováním.

Prohlašuji, že jsem tuto diplomovou práci vypracovala 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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne ..... Podpis autora

Název práce: Odhady parametrů založené na zaokrouhlených datech

Autor: Zuzana Dortová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: prof. RNDr. Jiří Anděl, DrSc.

Abstrakt: Tato práce pojednává o odhadech založených na zaokrouhlených datech. Práce popisuje odhady parametrů v časových řadách AR a MA a lineární regresi, uvádí různé metody odhadů na základě zaokrouhlených dat. Zaměřuje se zejména na model časové řady AR(1) a lineární regresi, kde teorii doplňuje simulacemi a porovnává metody na zaokrouhlených a nezaokrouhlených datech. Porovnání u lineární regrese navíc ilustruje na grafech.

Klíčová slova: zaokrouhlená data, časové řady, lineární regrese

Title: Estimates of parameters based on rounded data

Author: Zuzana Dortová

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Jiří Anděl, DrSc.

Abstract: This work discusses estimates based on rounded data. The work describes the estimates of parameters in time series AR and MA and in linear regression, the work presents different kinds of estimates based on rounded data. The work focuses on time series model AR(1) and linear regression, where simulations are added to theories and methods are compared on rounded and unrounded data. In addition, the comparison of linear regression is shown on the exemple of graph data.

Keywords: rounded data, time series, linear regression

Obsah

Kapitola 1

Úvod

Tato práce pojednává o odhadech parametrů založených na zaokrouhlených datech. Statistické modely obvykle předpokládají, že jsou k dispozici přesné hodnoty. Pokud měříme spojitou veličinu, data jsou zaokrouhlená, protože jsme limitováni přesností měření. V praxi občas zaokrouhlujeme i u diskrétních naměřených hodnot. Pokud bychom s daty počítali jako s přesnými a zaokrouhlení ignorovali, i při malých zaokrouhlovacích chybách může ve výsledku dojít k velkým odchylkám oproti výpočtům z přesných dat, zejména ve velkých souborech. Na problémy se zaokrouhlenými daty při tradičních statistických metodách poukázal jako první Sheppard (1898), dále např. Tricker (1990).

Vlivu zaokrouhlení se věnoval Lindley (1950), který pomocí maximálně věrohodné metody odvodil korekci pro zaokrouhlená data oproti přesným datům pro jednorozměrný i vícerozměrný parametr a odvodil některé jejich vlastnosti. Tallis (1967) zobecnil odvozené vzorce na mnohorozměrné normální rozdělení. Dempster a Rubin (1983) odvodili variantu Sheppardových korekcí pro nejmenší čtverce v lineárních modelech. Stam a Cogger (1993) se věnovali vyšetřování zaokrouhlování v gaussovských AR modelech, včetně nestacionárního AR(1) modelu a prováděli rozsáhlé simulace. Guo a Li (2012) odvodili korigované odhady pro MA modely.

Nejprve uvedeme nejběžnější metody odhadu parametrů. Poté uvedeme použití korekcí pro zaokrouhlená data. Nakonec ukážeme aplikaci na modelech časových řad a lineární regresi.

Kapitola 2

Metody odhadu parametrů

2.1 Momentová metoda

Použití momentové metody (dále též označována MM) vyžaduje, abychom znali typ rozdělení, ze kterého náhodný výběr pochází (normální, exponenciální, rovnoměrné, ...). Předpokládejme, že náhodný výběr $X_1, ..., X_n$ pochází z rozdělení s parametry $\theta_1, ..., \theta_k$.

Momentová metoda je založena na tom, že do vyjádření přesného 1. až k-tého momentu pomocí parametrů $\theta_1, ..., \theta_k$ dosadíme příslušné výběrové momenty napočítané z náhodného výběru. Tím obvykle dostaneme k rovnic pro k neznámých a jejich vyřešením získáme odhady parametrů.

Výběrové momenty jsou vyjádřeny následovně:

$$\text{obecný moment } k\text{-tého řádu: } m'_k = \frac{1}{n}\sum_{i=1}^{k} x_i^k,$$

$$\text{centrální moment } k\text{-tého řádu: } m_k = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^k.$$

Pokud odhadujeme parametry z vícerozměrného rozdělení, můžeme využít i smíšené momenty různých složek 2. a vyššího řádu, jako je např. kovariance.

Příklad: Mějme náhodný výběr $X_1, ..., X_n$ z normálního rozdělení $\mathrm{N}(\mu,\sigma_x^2)$. Normální rozdělení je určeno střední hodnotou a rozptylem. Odhady parametrů vypočítáme ze vzorců:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2.$$

Použití momentové metody je jednoduché. Její nevýhodou je, že počet rovnic je shora omezený v závislosti na tom, kolika parametry jsou určená rozdělení, ze kterých pochází náhodné veličiny. V některých případech se může stát, že počet parametrů, které budeme chtít odhadnout, bude větší než počet rovnic. Hlavní nevýhodou však bývá malá eficience momentové metody.

2.2 Metoda maximální věrohodnosti

2.2.1 Metoda maximální věrohodnosti pro přesná data

Použití metody maximální věrohodnosti (dále ji též budeme uvádět pod zkratkou MLE) předpokládá, že známe typ rozdělení, ze kterého náhodný výběr pochází. Uvažujme spojité rozdělení s hustotou $f(x, \theta)$, ke které existuje derivace podle θ, kde x je proměnná a θ je parametr, který chceme odhadnout. Mějme náhodný výběr $X_1, ..., X_n$ z rozdělení s hustotou $f(x, \theta)$.

Pokud je parametr θ jednorozměrný, jeho maximálně věrohodný odhad získáme zpravidla řešením rovnice

$$\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \ln f(X_i, \theta) = 0, \tag{2.1}$$

kde $\frac{\mathrm{d}}{\mathrm{d}\theta}$ je derivace podle θ a ln je přirozený logaritmus.

Pokud je parametr θ vícerozměrný (k-rozměrný), požadujeme, aby existovaly parciální derivace hustoty $f(x, \boldsymbol{\theta})$ podle jednotlivých složek parametru $\boldsymbol{\theta}$. Maximálně věrohodný odhad pro vícerozměrný parametr $\boldsymbol{\theta}$ získáme zpravidla řešením soustavy

$$\begin{aligned} &\sum_{i=1}^{n} \frac{\partial}{\partial\theta_1} \ln f(X_i, \theta_1, ..., \theta_k) = 0, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\sum_{i=1}^{n} \frac{\partial}{\partial\theta_k} \ln f(X_i, \theta_1, ..., \theta_k) = 0. \end{aligned} \tag{2.2}$$

2.2.2 Metoda maximální věrohodnosti pro jednorozměrný parametr ze zaokrouhlených dat

Nyní předpokládejme, že výběr je sdružen na intervalech délky h, přičemž 0 je v prostředku jednoho z těchto intervalů. Předpokládejme, že výběr pochází ze spojitého rozdělení s distribuční funkcí F a hustotou f.
Pak

$$\begin{aligned} p(nh, \theta) &= \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} f(x, \theta)\ \mathrm{d}x \\ &= F\left(\left[n + \frac{1}{2}\right] h, \theta\right) - F\left(\left[n - \frac{1}{2}\right] h, \theta\right) \end{aligned} \tag{2.3}$$

je pravděpodobnost, že hodnota náhodné veličiny X leží v intervalu délky h se středem nh, kde n je přirozené číslo. Hodnoty z jednoho takového intervalu se při zaokrouhlování na jednotky zaokrouhlí na stejné číslo.

Dále předpokládejme že existuje derivace hustoty $f(x, \theta)$ podle parametru θ.

Lindley (1950) uvádí, že po zaokrouhlení X má věrohodnostní rovnice pro jednorozměrný parametr θ tvar

$$\sum_{i=1}^{n} \frac{\mathrm{d}}{\mathrm{d}\theta} \ln p(n_i h, \theta) = 0, \tag{2.4}$$

kde X_i leží v intervalu

$$\left(\left[n_i - \frac{1}{2} \right] h \,;\, \left[n_i + \frac{1}{2} \right] h \right).$$

Vyjádření věrohodnostní rovnice pro zaokrouhlená data je poměrně intuitivní — po zaokrouhlení nahradíme ve vzorci 2.1 hustotu intergálem přes intervaly zaokrouhlení, protože na nich nejsme schopni zaokrouhlené hodnoty zpětně rozlišit. Výraz $p(nh, \theta)$ získáme integrací přes interval, ve kterém se náhodná veličina X zaokrouhlí na stejnou hodnotu. Tím $p(nh, \theta)$ vyjadřuje pravděpodobnost, že zaokrouhlená veličina leží v příslušném intervalu, po převodu ze spojitého případu na diskrétní s ní nahradíme původní spojitou hustotu.

Řešení věrohodnostní rovnice (vzorec 2.4) pro zaokrouhlená data označíme θ_1^*.

Chceme vypočítat hodnotu θ_1^*, ale máme vypočítanou pouze její aproximaci $\theta^*(n_1 h, ..., n_k h) = \theta_0^*$, získanou, jako by šlo o data nezaokrouhlená. Pokud odhad θ_0^* existuje a je jediný, ignoruje případné seskupování (intervalu hodnot je přiřazena jedna hodnota, např. při zaokrouhlování). Odhad neexistuje např. v situaci, kdy $p(n_i h, \theta) = 0$ pro některé $i \in \{1, 2, ..., n\}$. V takovém případě by logaritmus výrazu měl hodnotu $-\infty$.

Lindley (1950) vztah mezi θ_0^* a θ_1^* vyjadřuje pomocí aditivního členu Δ, jehož přičtením k θ_0^* dostane θ_1^*. Výsledný vztah $\theta_1^* = \theta_0^* + \Delta$ přirovnává k odhadu momentů a Sheppardovým korekcím.

Sheppardovy korekce jsou korekce vypočítané momentovou metodou z náhodně rozdělených dat, která jsou sdružena na intervalech stejné délky.

Dále uvedeme odvození odhadu pro Δ, které uvádí Lindley (1950) a v některých částech ho podrobněji rozvedeme (rozvoj pomocí Taylorova polynomu, odvození vyjádření ve tvaru polynomu s h a f, jeho logaritmování a derivaci). Předpokládejme, že hustota f má třetí derivaci podle x. Pro snazší úpravy rozvineme funkci f v bodě nh pomocí Taylorova polynomu.

Taylorův polynom uvádí Jarník (1974). Pro funkci f, která má $(n+1)$-ní derivaci při rozvíjení v bodě a, má rozvoj tvar

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + ... + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n+1}(x).$$

Nechť ξ leží na úsečce mezi body a a x. Zbytkový člen $R_{n+1}(x)$ po Taylorově

rozvoji lze vyjádřit jako Cauchyův zbytek ve tvaru

$$\frac{(x-a)^{n+1}}{n!} f^{(n+1)}(\xi)$$

nebo jako Lagrangeův zbytek ve tvaru

$$\frac{(x-\xi)^n (x-a)}{(n+1)!} f^{(n+1)}(\xi).$$

Při použití Taylorova rozvoje vyššího řádu bychom získali jemnější aproximaci, avšak požadovali bychom, aby hustota f měla derivace vyšších řádů. Pro hustotu f se třetí derivací rozvojem dostaneme

$$\begin{aligned} p\Big(nh, \theta\Big) &= \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} f(x,\theta)\ \mathrm{d}x \\ &= \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} \Big[f(nh,\theta) + f'(nh,\theta)(x-nh) \\ &\quad + \frac{f''(nh,\theta)}{2}(x-nh)^2 + Rh^3 \Big]\ \mathrm{d}x, \end{aligned}$$

kde R je člen obsahující třetí derivaci funkce f. Při použití Cauchyova zbytku uvedého výše platí $R = \frac{1}{2} f'''(\xi)(x-nh)^3$.

V následující části budeme značit $f := f\,(nh, \theta)$. Nejprve budeme integrovat výrazy, které v integrálu neobsahují argument x. Poté zintegrujeme výrazy, které argument x obsahují, a výsledné členy upravíme, abychom získali zintegrované výrazy a z nich odvodili korekci pro maximálně věrohodný odhad ze zaokrouhlených dat.

Postupně dostaneme

$$\begin{aligned} \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} f\,(x,\theta)\ \mathrm{d}x &= \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} \left[f + f'\,(x-nh) + \frac{f''}{2}\,(x-nh)^2 + Rh^3 \right]\ \mathrm{d}x \\ &= hf - nh^2 f' + \int_{(n-\frac{1}{2})h}^{(n+\frac{1}{2})h} \left[f'(x) + \frac{f''}{2}\left(x^2 - 2xnh + n^2h^2\right) \right. \\ &\quad \left. + Rh^3 \right]\ \mathrm{d}x \end{aligned}$$

$$
\begin{aligned}
&= hf - nh^2 f' + n^2 h^3 \frac{f''}{2} + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} \left[f'x + \frac{f''}{2}\left(x^2 - 2xnh\right) \right. \\
&\quad \left. + Rh^3 \right] \mathrm{d}x \\
&= hf - nh^2 f' + n^2 h^3 \frac{f''}{2} + f' \left[\frac{x^2}{2} \right]_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} - 2nh \frac{f''}{2} \left[\frac{x^2}{2} \right]_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} \\
&\quad + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} \left(\frac{f''}{2} x^2 + Rh^3 \right) \, \mathrm{d}x \\
&= hf - nh^2 f' + n^2 h^3 \frac{f''}{2} + f' \left[\frac{\left(\left(n+\frac{1}{2}\right)h\right)^2}{2} - \frac{\left(\left(n-\frac{1}{2}\right)h\right)^2}{2} \right] \\
&\quad - 2nh \frac{f''}{2} \left[\frac{\left(\left(n+\frac{1}{2}\right)h\right)^2}{2} - \frac{\left(\left(n-\frac{1}{2}\right)h\right)^2}{2} \right] + \frac{f''}{2} \left[\frac{x^3}{3} \right]_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} \\
&\quad + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \, \mathrm{d}x \\
&= hf - nh^2 f' + n^2 h^3 \frac{f''}{2} + f' \frac{h^2}{2} \left[\left(n+\frac{1}{2}\right)^2 - \left(n-\frac{1}{2}\right)^2 \right] \\
&\quad - nhf'' \frac{h^2}{2} \left[\left(n+\frac{1}{2}\right)^2 - \left(n-\frac{1}{2}\right)^2 \right] + \frac{f''}{2} \left[\frac{\left(\left(n+\frac{1}{2}\right)h\right)^3}{3} \right. \\
&\quad \left. - \frac{\left(\left(n-\frac{1}{2}\right)h\right)^3}{3} \right] + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \, \mathrm{d}x \\
&= hf - nh^2 f' + n^2 h^3 \frac{f''}{2} + f' \frac{h^2}{2} \left[\left(n+\frac{1}{2}\right)^2 - \left(n-\frac{1}{2}\right)^2 \right] \\
&\quad - nhf'' \frac{h^2}{2} \left[\left(n+\frac{1}{2}\right)^2 - \left(n-\frac{1}{2}\right)^2 \right] + \frac{f''}{2} \frac{h^3}{3} \left[\left(n+\frac{1}{2}\right)^3 \right. \\
&\quad \left. - \left(n-\frac{1}{2}\right)^3 \right] + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \, \mathrm{d}x.
\end{aligned}
$$

Jelikož

$$
\begin{aligned}
\left[\left(n+\frac{1}{2}\right)^2 - \left(n-\frac{1}{2}\right)^2 \right] &= 2n, \\
\left[\left(n+\frac{1}{2}\right)^3 - \left(n-\frac{1}{2}\right)^3 \right] &= 3n^2 + \frac{1}{4},
\end{aligned}
$$

po dosazení dostáváme

$$\begin{aligned}
p\left(nh,\theta\right) &= hf - nh^2f' + n^2h^3\frac{f''}{2} + f'\frac{h^2}{2}2n - nhf''\frac{h^2}{2}2n + \frac{f''}{2}\frac{h^3}{3}\left(3n^2+\frac{1}{4}\right) \\
&\quad + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \,\mathrm{d}x \\
&= hf - nh^2f' + n^2h^3\frac{f''}{2} + f'nh^2 - f''n^2h^3 + \frac{f''}{2}\frac{h^3}{3}\left(3n^2+\frac{1}{4}\right) \\
&\quad + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \,\mathrm{d}x \\
&= hf + f''\left(n^2h^3\frac{1}{2} - n^2h^3 + \frac{h^3}{2}n^2\right) + \frac{f''h^3}{24} + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \,\mathrm{d}x \\
&= hf + \frac{f''h^3}{24} + \int_{\left(n-\frac{1}{2}\right)h}^{\left(n+\frac{1}{2}\right)h} Rh^3 \,\mathrm{d}x \\
&= hf + \frac{f''h^3}{24} + R'h^4\,,
\end{aligned}$$

kde R' je zbytkový člen obsahující derivaci třetího řádu funkce $f(nh,\theta)$.

Za předpokladu, že $f(nh,\theta)$ není nula (v takovém případě by byl její logaritmus $-\infty$ a nemohli bychom použít maximálně věrohodný odhad), dostáváme

$$\begin{aligned}
\ln p\left(nh,\theta\right) &= \ln\left(hf + \frac{f''h^3}{24} + R'h^4\right) \\
&= \ln\left[hf\left(1 + \frac{f''h^2}{24f} + \frac{R'h^3}{f}\right)\right] \\
&= \ln hf + \ln\left(1 + \frac{f''h^2}{24f} + \frac{R'h^3}{f}\right).
\end{aligned}$$

Derivováním podle θ dostáváme

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\ln p\left(nh,\theta\right) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln hf + \frac{\mathrm{d}}{\mathrm{d}\theta}\ln\left(1 + \frac{f''h^2}{24f} + \frac{R'h^3}{f}\right) \\
&= \frac{1}{hf}h\frac{\mathrm{d}}{\mathrm{d}\theta}f + \frac{1}{\left(1 + \frac{f''h^2}{24f} + \frac{R'h^3}{f}\right)}\frac{\mathrm{d}}{\mathrm{d}\theta}\left(1 + \frac{f''h^2}{24f} + \frac{R'h^3}{f}\right) \\
&= \frac{1}{f}\frac{\mathrm{d}}{\mathrm{d}\theta}f + \frac{1}{\left(\frac{24f+f''h^2+24R'h^3}{24f}\right)}\left(\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f} + R'h^3\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{24f}{(24f + f''h^2 + 24R'h^3)}\left(\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f} + R'h^3\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}\right)\\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{24f}{(24f + f''h^2 + 24R'h^3)}\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f}\\
&\quad + \frac{24fR'h^3}{(24f + f''h^2 + 24R'h^3)}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}\\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{24f}{(24f + f''h^2 + 24R'h^3)}\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f}\\
&\quad + \frac{24fR'h^3}{(24f + f''h^2 + 24R'h^3)}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}\\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{(24f + f''h^2 + 24R'h^3 - f''h^2 - 24R'h^3)}{(24f + f''h^2 + 24R'h^3)}\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f}\\
&\quad + \frac{24fR'h^3}{(24f + f''h^2 + 24R'h^3)}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}\\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f} + \frac{(-f''h^2 - 24R'h^3)}{(24f + f''h^2 + 24R'h^3)}\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f}\\
&\quad + \frac{24fR'h^3}{(24f + f''h^2 + 24R'h^3)}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{f}.
\end{aligned}
$$

Jelikož poslední dva členy obsahují třetí mocninu h, můžeme je shrnout pod výraz $O(h^3)$. Shrnutí výrazu pod $O(h^3)$ znamená, že pro malé hodnoty je výraz až na multiplikační konstantu menší než h^3. Matematicky zapsáno, $f(x)$ je $O(g(x))$, právě když existuje $C > 0, x_0$, že $\forall x < x_0$ platí $|f(x)| \leq |Cg(x)|$. (Funkce f v této definici není konkrétní funkcí f v odvození korekce Δ, jedná se o obecný zápis definice.)

Z výše uvedeného dostaneme

$$
\frac{\mathrm{d}}{\mathrm{d}\theta}\ln p(nh, \theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f} + O(h^3).
$$

Lindley (1950) navrhuje řešit rovnici

$$
\sum_{i=1}^{k}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f + \frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{f''}{f}\right] = 0
$$

Newtonovou metodou (někdy též nazývána Newton-Rawsonova metoda) s počáteční hodnotou parametru θ_0^*, vypočítaného z rovnice (vzorec 2.1), jako by šlo o nezaokrouhlená data s prvotní aproximací $\sum_{i=1}^{k}\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f(n_i h, \theta_0^*) = 0$.

Newtonova metoda je iterační metoda sloužící k aproximaci kořenů rovnic reálných funkcí. Je založena na lineární aproximaci pomocí tečny. Pro její použití

k nalezení řešení x rovnice $f(x) = 0$ potřebujeme, aby funkce f měla první derivaci. Zvolíme počáteční bod x_0. Iterační kroky mají tvar $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$. (I v tomto případě popisu Newtonovy metody se jedná o obecnou funkci f, nikoliv o konkrétní z odvozování korekce Δ.)

Jako výslednou korekci Lindley (1950) uvádí

$$\Delta = -\frac{h^2}{24}\left[\frac{\sum_{i=1}^{k}\frac{\mathrm{d}}{\mathrm{d}\theta}(\frac{f''}{f})_{\theta_0^*}}{\sum_{i=1}^{k}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\ln f)_{\theta_0^*}}\right] + O(h^3).$$

Tallis (1967) uvádí vztah odhadu na zaokrouhlených datech při igonorování zaokrouhlení s korekcí v jiném tvaru $\theta_1^* = \theta_0^* + \Delta$, kde $\Delta = V(\theta_0^*)e(\theta_0^*)$, $[V(\theta_0^*)]^{-1} = -E(\frac{\mathrm{d}^2 \ln f(x,\theta_0^*)}{\mathrm{d}\theta^2})$, $e(\theta_0^*) = E(\frac{\mathrm{d}[\frac{hf''}{24f}]}{\mathrm{d}\theta})$, kde $f'' = \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}$.

Při ignorování členu $O(h^3)$ snadno převedeme jeden tvar korekce na druhý.

Vyjdeme z Lindleyho tvaru

$$\Delta = -\frac{h^2}{24}\left[\frac{\sum_{i=1}^{k}\frac{\mathrm{d}}{\mathrm{d}\theta}(\frac{f''}{f})_{\theta_0^*}}{\sum_{i=1}^{k}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\ln f)_{\theta_0^*}}\right].$$

Zlomek $-\frac{h^2}{24}$ přesuneme do čitatele za sumu a výraz rozšíříme $\frac{1}{k}$. Dostaneme

$$\left[\frac{1}{k}\frac{\sum_{i=1}^{k}-\frac{h^2}{24}\frac{\mathrm{d}}{\mathrm{d}\theta}(\frac{f''}{f})_{\theta_0^*}}{\frac{1}{k}\sum_{i=1}^{k}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\ln f)_{\theta_0^*}}\right].$$

Průměr přes k sčítanců vyjadřuje střední hodnotu. Zlomek $\frac{h^2}{24}$ můžeme vložit do derivace, protože nezávisí na θ. Dostaneme

$$\left[-\frac{E\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{h^2}{24}(\frac{f''}{f})_{\theta_0^*}}{E\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}(\ln f)_{\theta_0^*}}\right].$$

Což lze vyjádřit jako $\frac{e(\theta_0^*)}{V(\theta_0^*)}$. Z výše uvedeného vyjádření uvažujeme, že když platí $[V(\theta_0^*)]^{-1} = -E(\frac{\mathrm{d}^2 \ln f(x,\theta_0^*)}{\mathrm{d}\theta^2})$, tak $V(\theta_0^*) = -E(\frac{\mathrm{d}^2 \ln f(x,\theta_0^*)}{\mathrm{d}\theta^2})^{-1}$.

2.2.3 Metoda maximální věrohodnosti pro vícerozměrný parametr ze zaokrouhlených dat

Kromě odhadu jednorozměrného parametru metodou maximální věrohodnosti na základě zaokrouhlených dat uvádí Lindley (1950) i variantu pro dvourozměrný parametr. Pro dvourozměrný parametr $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$ a korekce $\theta_{1,1} = \theta_{1,0} +$

$\Delta_1, \theta_{2,1} = \theta_{2,0} + \Delta_2$, kde $(\Delta_1, \Delta_2)^T$ jsou korekce, $(\theta_{1,1}, \theta_{2,1})^T$ jsou skutečné hodnoty parametru a $(\theta_{1,0}, \theta_{2,0})^T$ jsou hodnoty odhadnuté z rovnic, jako by šlo o přesná data (ignorující zaokrouhlení).

Lindley (1950) uvádí, že korekce lze získat řešením soustavy rovnic

$$\Delta_1 \sum \frac{\partial^2}{\partial \theta_{1,1}^2} \ln f + \Delta_2 \sum \frac{\partial^2}{\partial \theta_{1,1} \partial \theta_{2,1}} \ln f = -\frac{h^2}{24} \sum \frac{\partial}{\theta_{1,1}} (\frac{f''}{f}),$$
$$\Delta_1 \sum \frac{\partial^2}{\partial \theta_{1,1} \partial \theta_{2,1}} \ln f + \Delta_2 \sum \frac{\partial^2}{\partial \theta_{2,1}^2} \ln f = -\frac{h^2}{24} \sum \frac{\partial}{\theta_{2,1}} (\frac{f''}{f}).$$

Tallis (1967) uvádí rozšíření na vícerozměrný případ pro k-dimenzionální funkci s s-rozměrným parametrem $\boldsymbol{\theta}$. Pro každý rozměr k-rozměrného vektoru $\boldsymbol{X}$ uvažuje dělicí interval h_i, $i = 1, ..., k$. Ve vícerozměrném případě platí $\boldsymbol{\theta}_1^* = \boldsymbol{\theta}_0^* + \boldsymbol{\Delta}$,

kde $\boldsymbol{\Delta} = \boldsymbol{V}(\theta_0^*)\boldsymbol{e}(\theta_0^*)$, a $[\boldsymbol{V}(\boldsymbol{\theta}_0^*)]^{-1} = \boldsymbol{I}(\boldsymbol{\theta}_0^*) - E(\frac{\partial^2 \ln f(\boldsymbol{x}, \boldsymbol{\theta}_0^*)}{\partial \theta_i \partial \theta_j})$,

$\boldsymbol{e}(\boldsymbol{\theta}_0^*) = (e_1(\boldsymbol{\theta}_0^*), ..., e_s(\boldsymbol{\theta}_0^*))^T$,

$e_i(\boldsymbol{\theta}_0^*) = E\left(\frac{\partial [\sum_{j=1}^{k} \frac{h_j^2 f_{jj}(\boldsymbol{x}, \boldsymbol{\theta}_0^*)}{24 f(\boldsymbol{x}, \boldsymbol{\theta}_0^*)}]}{\partial \theta_i} \right)$, kde $f_{jj}(\boldsymbol{x}, \boldsymbol{\theta}_0^*) = \frac{\partial^2 f}{\partial x_j^2}$.

2.3 Vlastnosti metody maximální věrohonosti pro zaokrouhlená data

2.3.1 Porovnání metody maximální věrohodnosti a metody momentů pro zaokrouhlená data

Lindley (1950) uvádí nejen příklady, kdy korekce získaná metodou maximální věrohodnosti a momentovou metodou (Sheppardovy korekce) jsou stejné, ale i příklad, kdy jsou výsledky odlišné. Lindley (1950) uvádí i vysvětlení — korekce získané metodou maximální věrohodnosti konvergují k Sheppardovým korekcím, odlišné výsledky dostáváme, když ještě není dosaženo limity.

2.3.2 Ztráta eficience způsobená zaokrouhlením

Lindley (1950) uvádí, že maximálně věrohodný odhad má v limitě normální rozdělení s rozptylem $[-nE(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln f)]^{-1}$, což je nejvíce eficientní případ.

Lindley (1950) dále uvádí i následující výpočet ztráty eficience. Optimální eficienci vyjádříme jako zlomek $\frac{\mathrm{var}(\hat{\theta})}{\mathrm{var}(\theta_1)}$. Pak

$$E\left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln p(nh,\theta)\right) = \sum_{n=-\infty}^{\infty}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln p(nh,\theta)$$
$$= \sum_{n=-\infty}^{\infty}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f + \frac{h^2}{24}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\left(\frac{f''}{f}\right)\right]\left[hf + \frac{h^2}{24}f'\right] + O(h^4).$$

Při použití Euler-MacLaurinova vzorce je výraz roven

$$\int_{-\infty}^{\infty} f\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f\,\mathrm{d}x + \frac{h^2}{24}\int_{-\infty}^{\infty}\left[f\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\frac{f''}{f} + f''\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f\right]\mathrm{d}x,$$

kde eficience je

$$\frac{\operatorname{var}\hat{\theta}}{\operatorname{var}\theta_1} = 1 + \frac{h^2}{24}\frac{E\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\frac{f''}{f} + \frac{f''}{f}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f\right]}{E\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f\right]} + O(h^4).$$

2.4 Metoda nejmenších čtverců

Metoda nejmenších čtverců (dále též označována MNČ) se používá zejména při odhadování parametrů pomocí polynomiální regrese. Mějme veličinu X a závislou veličinu Y, o které víme, že je na X závislá vztahem $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + \varepsilon$, kde ε jsou nezávislé stejně rozdělené chyby s nulovou střední hodnotou a kladným rozptylem σ_ε^2. Obvykle známe i stupeň polynomu p.

Napozorovány máme veličiny $Y_1, ..., Y_n$ a $X_1, ..., X_n$. Vektor veličin $(Y_1, ..., Y_n)^T$ označíme $\boldsymbol{Y}$. Vektor parametrů $(\beta_0, ...\beta_p)^T$ označíme $\boldsymbol{\beta}$. Matici

$$\begin{pmatrix} 1 & X_1 & ... & X_1^p \\ 1 & X_2 & ... & X_2^p \\ ... & ... & ... & ... \\ 1 & X_n & ... & X_n^p \end{pmatrix}$$

o rozměrech $(p+1) \times n$ označme jako $\boldsymbol{Z}$.

Metoda nejmenších čtverců spočívá v minimalizování výrazu

$$(\boldsymbol{Y} - \boldsymbol{Z\beta})^T(\boldsymbol{Y} - \boldsymbol{Z\beta}),$$

což je součet obsahu čtverců, které se svými rohy dotýkají polynomiální funkce a jejichž strany jsou rovnoběžné s osami x a y. Při použití metody nejmenších

čtverců získáme odhad parametru $\boldsymbol{\beta}$ pomocí vzorce $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Y}$. Důkaz uvádí Anděl (2007b). Při metodě nejmenších čtverců se předpokládá, že jsou k dispozici přesná data. Vliv zaokrouhlení dat na odhady získané metodou nejmenších čtverců si ukážeme pomocí simulací ve 4. kapitole.

Kapitola 3

Aplikace na modely časových řad

3.1 Model MA

Bai a kol. (2009) uvádějí odhady parametrů v MA modelu (model klouzavých průměrů) na základě zaokrouhlených dat.

Mějme MA model:

$$X_t = c + \sum_{l=0}^{p} \varepsilon_{t-l}\phi_l,$$

kde ε_l jsou vzájemně nezávislé náhodné chyby pocházející z rozdělení $\mathrm{N}(0, \sigma^2), l = 0, ..., n$ a $\phi_0 = 1$.

Nechť $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)^T$. V uvedeném případě má vektor $(X_1, ..., X_n)^T$ normální rozdělení $N(c\mathbf{1}_n, \boldsymbol{\Sigma}_{n\times n})$, kde $\mathbf{1}_n$ je sloupcový vektor jedniček a $\boldsymbol{\Sigma}_{n\times n} = (\sigma_{ij})_{n\times n}$, $\sigma_{ij} = \gamma_{|i-j|}$,

$$\boldsymbol{\Sigma}_{n\times n} = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \cdots & \cdots & \cdots & \cdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix},$$

kde

$$\gamma_i = \begin{cases} \sigma^2(\phi_i + \phi_{i+1}\phi_1 + \phi_{i+2}\phi_2 + \cdots + \phi_p\phi_{p-i}), & i = 0, 1, ..., p, \\ 0 & \text{jinak.} \end{cases}$$

Pozorovat můžeme pouze zaokrouhlená data $\tilde{X}_1, ..., \tilde{X}_n$, kde $\tilde{X}_i$ je zaokrouhlená veličina X_i pro $i = 1, ..., n$. Bez újmy na obecnosti můžeme předpokládat, že data jsou zaokrouhlena na celá čísla.

Nechť $\mathbf{i} = (i_1, ..., i_{p+1})$ a i_j, $j = 1, ..., p+1$, jsou celá čísla. Pak $\mathbf{i}$ je vyjádření kombinace $p+1$ hodnot (ne nutně různých), kterých nabývá prvních $(p+1)$ zaokrouhlených hodnot náhodného výběru.

Pravděpodobnost, že pro určitou kombinaci nabudou zaokrouhlená data takových hodnot, označíme $p_{\mathbf{i}}$. Tedy

$$p_{\mathbf{i}} = P(i_j - 0,5 \leq X_j < i_j + 0,5), \qquad j = 1, ..., p+1. \tag{3.1}$$

Nechť $A_{\mathbf{i}}$ je obdélník

$$A_{\mathbf{i}} = \prod_{j=1}^{p+1} [i_j - 0,5; i_j + 0,5).$$

Jedná se o obdélník (od 3 rozměrů se jedná o kvádr) v prostoru, se středem v bodě $(\tilde{X}_1, ..., \tilde{X}_{p+1})$. Toto značení uvádíme, protože Bai a kol. (2009) jej využívají při důkazu vlastností odhadu.

Příklad: Mějme $p = 1$ a MA model s $c = 0$, $\phi_1 = 0,5$ a ε_t vzájemně nezávislé z $R(-1, 1)$; takto označujeme rovnoměrné rozdělení na intervalu $(-1; 1)$.

Pak $\tilde{X}_i$ po zaokrouhlení na jednotky může nabýt hodnot -1, 0 a 1. Jelikož $p = 1$, označuje index **i** dvojici. Index **i** může nabýt devíti různých hodnot $(-1, -1)$, $(-1, 0)$, $(-1, 1)$, $(0, -1)$, $(0, 0)$, $(0, 1)$, $(1, -1)$, $(1, 0)$ a $(1, 1)$. Pravděpodobnosti, že data nabudou dvojice hodnot, můžeme získat pomocí výpočtů nebo z geometrického zobrazení (z tohoto důvodu používáme v příkladu rovnoměrné a nikoliv normální rozdělení - při rovnoměrném rozdělením jsou pravděpodobnosti úměrné plochám v grafu). Příklad není převzat, uvádíme ho pro lepší názornost pochopení obdélníků, které zavádí Bai a kol.(2009).

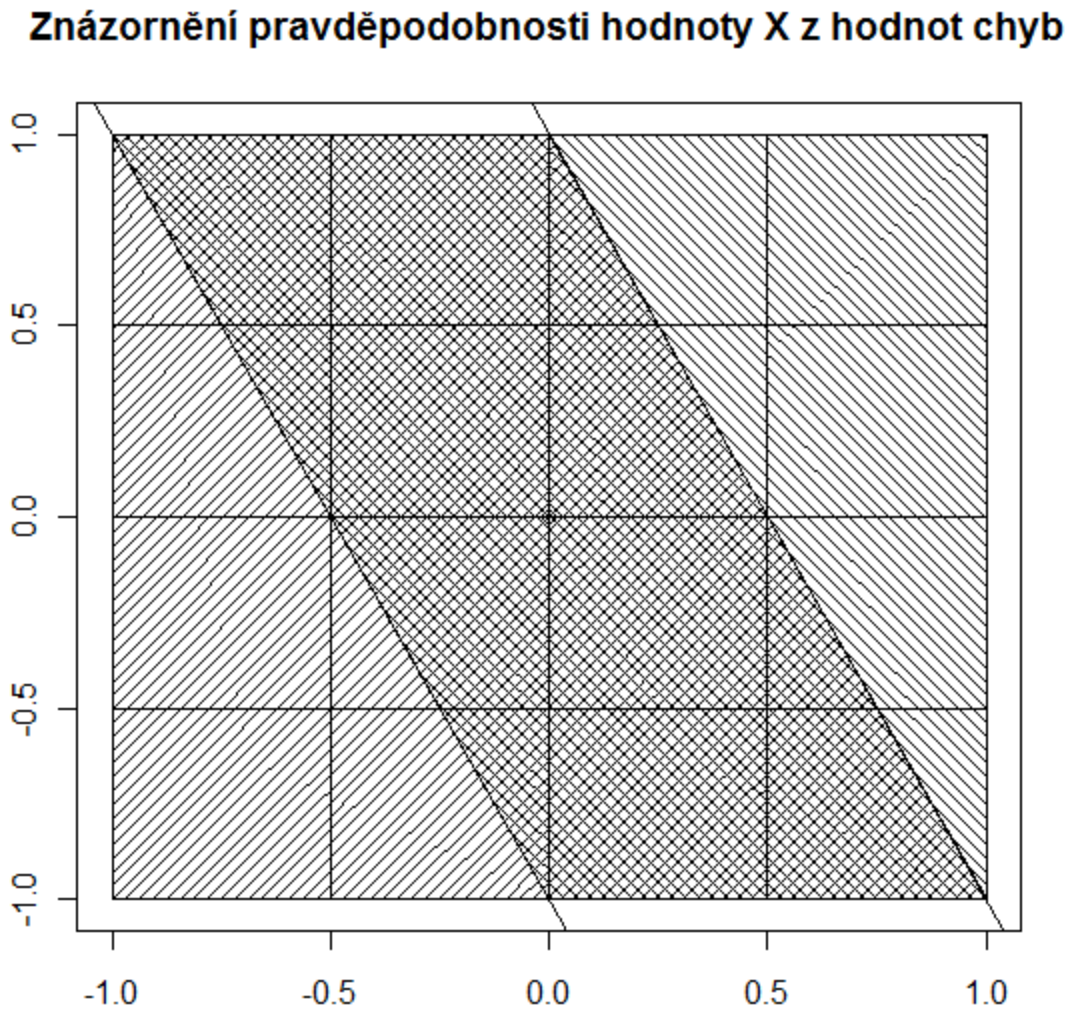
Nejprve získáme pravděpodobnosti první hodnoty na základě kombinací dvojic $\varepsilon_{t-1}, \varepsilon_t$. Pomocí grafického znázornění, kdy na osy naneseme hodnoty ε_{t-1} a ε_t, dostaneme pravděpodobnosti první hodnoty $(\tilde{X}_1)$ z dvojice $(\tilde{X}_1, \tilde{X}_2)$. Jelikož ε_t pocházejí z rovnoměrného rozdělení, je pravděpodobnost nabytí hodnoty úměrná ploše v obdélníku $[-1, 1] \times [-1, 1]$. Pravděpodobnost nabytí hodnoty nula je $\frac{1}{2}$ (čtverečkovaná plocha na obrázku 3.1), pro hodnoty 1 a -1 je to $\frac{1}{4}$ (každá ze šrafovaných ploch na obrázku 3.1).

Pokud má $\tilde{X}_t$ hodnotu -1, pak ε_t musí ležet v intervalu $(-1, 0]$. V takovém případě pravděpodobnost nabytí druhé hodnoty z dvojice $(\tilde{X}_{t+1})$ určíme pouze na základě obdélníků pod osou x — tyto hodnoty jsou popořadě $\frac{3}{8}$, $\frac{4}{8}$, $\frac{1}{8}$, pro -1,0 a 1. Pro $\tilde{X}_t$ s hodnotou 1 musí ε_t ležet v intervalu $[0, 1)$. Pravděpodobnosti nabytí hodnoty X_{t+1} určíme díky symetrii rozdělení ε_{t-1} obdobně jako pro ε_{t-1} o hodnotě -1. Pokud $\tilde{X}_t$ má hodnotu 0, může ε_t ležet v celém intervalu $[-1, 1)$. Určení pravděpodobnosti nabytí hodnot $\tilde{X}_{t+1}$ je stejné jako pro určení pravděpodobnosti nabytí hodnot $\tilde{X}_t$.

Pro hodnoty -1 a 1 zjistíme podmíněné pravděpodobnosti pro ε_{t+1} opět z grafického znázornění. Vše je vidět na obrázku 3.1, kde na ose x máme znázorněnu hodnotu ε_i a na ose y hodnotu ε_{i+1}.

Pro první hodnotu -1 jsou pravděpodobnosti druhé hodnoty popořadě $\frac{3}{8}$, $\frac{4}{8}$, $\frac{1}{8}$, pro hodnotu 1 ze symetrie $\frac{1}{8}$, $\frac{4}{8}$ a $\frac{3}{8}$.

Obrázek 3.1: Grafické znázornění pravděpodobností pro kombinace ε_i (osa x) a ε_{i+1} (osa y) k příkladu MA(1) s $\varepsilon \in \mathrm{R}(-1,1)$ modelu



Z tohoto grafu odvodíme pravděpodobnosti jednotlivých kombinací $p_\mathbf{i}$ pro $c = 0$, $\phi_1 = 0,5$ a $\varepsilon_t \in \mathrm{R}(-1{,}1)$. Rovnoměrné rozdělení chyb je v tomto příkladu použito kvůli snadnější ilustraci, přestože v modelech časových řad předpokládáme chyby z normálního rozdělení.

Pronásobením pak získáme hodnoty $p_{\mathbf{i}}$:

$$\begin{aligned}
p_{(-1,1)} &= \frac{1}{4} \times \frac{3}{8} = \frac{3}{32}, \\
p_{(-1,0)} &= \frac{1}{4} \times \frac{4}{8} = \frac{4}{32}, \\
p_{(-1,1)} &= \frac{1}{4} \times \frac{1}{8} = \frac{1}{32}, \\
p_{(0,-1)} &= \frac{1}{2} \times \frac{1}{4} = \frac{4}{32}, \\
p_{(0,0)} &= \frac{1}{2} \times \frac{1}{2} = \frac{8}{32}, \\
p_{(0,1)} &= \frac{1}{2} \times \frac{1}{4} = \frac{4}{32}, \\
p_{(1,-1)} &= \frac{1}{4} \times \frac{1}{8} = \frac{1}{32}, \\
p_{(1,0)} &= \frac{1}{4} \times \frac{4}{8} = \frac{4}{32}, \\
p_{(1,1)} &= \frac{1}{4} \times \frac{3}{8} = \frac{3}{32}.
\end{aligned}$$

Obdélníky $A_{\mathbf{i}}$ mají středy v bodech $(-1,-1)$, $(-1,0)$, $(-1,1)$, $(0,-1)$, $(0,0)$, $(0,1)$, $(1,-1)$, $(1,0)$ a $(1,1)$.

Dále popíšeme aproximaci maximálně věrohodného odhadu, jak ho uvádí Bai a kol.(2009). Označme náhodný výběr $(X_1, ..., X_n)$ jako sekvenci $\boldsymbol{X}$ a jeho zaokrouhlení $(\tilde{X}_1, ..., \tilde{X}_n)$ jako sekvenci $\tilde{\boldsymbol{X}}$. Pak sekvence veličin $\boldsymbol{X}$ a $\tilde{\boldsymbol{X}}$ jsou p-závislé; veličiny (X_i,X_j), $i, j \in N_0$ jsou závislé pouze tehdy, pokud $|i - j| \leq p$. Můžeme proto data rozdělit na p podsekvencí nezávislých stejně rozdělených náhodných veličin a odhadovat parametry na základě každé podsekvence.

Položme $m = \frac{n-1}{p+1}$. Definujme $p + 1$ podsekvencí následujícím způsobem:

1) $\tilde{X}_1, ..., \tilde{X}_{p+1}, \tilde{X}_{2p+2}, ..., \tilde{X}_{3p+2}$ ... $\tilde{X}_{(m-1)(p+1)+1}, ..., \tilde{X}_{m(p+1)}$,

2) $\tilde{X}_2, ..., \tilde{X}_{p+2}, \tilde{X}_{2p+3}, ..., \tilde{X}_{3p+3}$ ... $\tilde{X}_{(m-1)(p+1)+2}, ..., \tilde{X}_{m(p+1)+1}$,

... .

p+1) $\tilde{X}_{p+1}, ..., \tilde{X}_{2p+1}, \tilde{X}_{3p+2}, ..., \tilde{X}_{4p+2}$ ... $\tilde{X}_{m(p+1)}, ..., \tilde{X}_{m(p+1)+p}$.

Odhad ze zaokrouhlených dat $\tilde{X}_1, ..., \tilde{X}_n$ dostaneme následovně:

1) V sekvenci $(\tilde{X}_1 ... \tilde{X}_{p+1})$, $(\tilde{X}_{2p+2} ... \tilde{X}_{3p+2})$, ..., $(\tilde{X}_{(m-1)(p+1)+1} ... \tilde{X}_{m(p+1)})$ máme m nezávislých vektorů o $p + 1$ složkách, které mají stejné rozdělení. Četnosti $(p + 1)$-tic $\boldsymbol{i}$ označíme $n_{\boldsymbol{i}}$.

Na základě takto získaných hodnot vypočítáme maximálně věrohodné odhady parametrů $(c, \boldsymbol{\phi}, \sigma^2)$ maximalizací výrazu $\sum_{\mathbf{i}} n_{\mathbf{i}} \ln p_{\mathbf{i}}$. Tím získáme maximálně věrohodný odhad pravděpodobností jednotlivých $(p + 1)$-tic hodnot $p_{\mathbf{i}}$, z nich potom odvodíme maximálně věrohodné odhady parametrů. Výsledný maximálně věrohodný odhad označíme $(\hat{c_1}, \hat{\boldsymbol{\phi}}_1, \hat{\sigma_1^2})$.

2) Pro podposloupnosti uvedené výše v tabulce na řádcích $2, ..., p+1$ zkonstruujeme obdobným postupem maximálně věrohodné odhady $(\hat{c_j}, \hat{\boldsymbol{\phi}}_j, \hat{\sigma_j^2})$, $j = 2, ..., p+1$, parametrů $(c, \boldsymbol{\phi}, \sigma^2)$.

3) Výsledné odhady parametrů získáme zprůměrováním dílčích odhadů:

$$\hat{c} = \sum_{j=1}^{p+1} \frac{\hat{c_j}}{p+1},$$

$$\hat{\boldsymbol{\phi}} = \sum_{j=1}^{p+1} \frac{\hat{\boldsymbol{\phi}}_j}{p+1},$$

$$\hat{\sigma^2} = \sum_{j=1}^{p+1} \frac{\hat{\sigma_j^2}}{p+1},$$

čímž dostaneme aproximaci maximálně věrohodného odhadu $(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma^2})$.

Bai a kol. (2009) dále uvádějí věty související s vlastnostmi aproximace maximálně věrohodného odhadu, kterou označujeme zkratkou AMLE.

Věta 3.1 *Slabě stacionární gaussovský proces* $\{X_t, t \in T\}, T \subset R$, *kde* R *je množina reálných čísel, je striktně stacionární.*

Důkaz uvádí Prášková(2004) jako větu 3.1 .

Věta 3.2 *Posloupnost* $\{X_t, t \in Z\}$ *klouzavých součtů řádu* n *definovaná vztahem* $X_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + ... + b_n\varepsilon_{t-n}, t \in Z$, *kde* $\{\varepsilon_t, t \in Z\}$ *je bílý šum se střední hodnotou 0 a rozptylem* σ^2 *a* $b_0, ..., b_n$ *jsou reálné nebo komplexní konstanty,* $b_0 \neq 0, b_n \neq 0$ *je centrovaná a slabě stacionární.*

Toto tvrzení je částí věty 5.1., kterou dokazuje Prášková (2004).

Bai a kol. (2009) uvádějí a dokazují větu 3.3. Zde je navíc v důkazu zdůvodněna striktní stacionarita MA(p) řady odkazem na výše uvedené věty ze skript Prášková(2004).

Věta 3.3 *Aproximace MLE* $(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma^2})$ *uvedená výše na základě zaokrouhlených dat* $\tilde{X_1}, ..., \tilde{X_n}$ *je konzistentní.*

Důkaz: Gaussovská MA(p) časová řada je p-závislá a striktně stacionární, takže skupiny zaokrouhlených pozorování $(\tilde{X_1}...\tilde{X}_{p+1}), (\tilde{X}_{2p+2}...\tilde{X}_{3p+2}), ..., (\tilde{X}_{(m-1)(p+1)+1}...\tilde{X}_{m(p+1)})$ jsou nezávislé a stejně rozdělené. První vlastnost, p-závislost, plyne z vyjádření MA(p) řady, kde X_t závisí pouze na posledních $p+1$ členech $\varepsilon_{t-p}, ..., \varepsilon_t$. Striktní stacionarita plyne z vět 3.1 a 3.2. Potom odhad $(\hat{c_1}, \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2)$ je silně konzistentní. Obdobně jsou silně konzistentní i ostatní dílčí odhady parametrů $(\hat{c_j}, \hat{\boldsymbol{\phi}}_j, \hat{\sigma}_j^2), j = 2, ..., p+1$.

Pak pro $n \to \infty$ je odhad $(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma^2})$ výše uvedeným způsobem na základě $\tilde{X_1}, ..., \tilde{X_n}$ silně konzistentní. □

Následující větu 3.4 i s důkazem uvádí Bai a kol. (2009), dokazují v ní důležité vlastnosti AMLE odhadu. V této práci podrobněji uvádíme odvození přibližné hodnoty $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$.

Věta 3.4 *Nechť*

$$\begin{aligned}
\boldsymbol{\theta} &= (c, \boldsymbol{\phi}, \sigma^2)^T, \\
\frac{\mathrm{d}p_{\mathbf{i}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} &= \left(\frac{\partial p_{\mathbf{i}}(\boldsymbol{\theta})}{\partial \theta_1}, ..., \frac{\partial p_{\mathbf{i}}(\boldsymbol{\theta})}{\partial \theta_{p+2}} \right), \\
I(\boldsymbol{\theta}) &= \sum_{i=-\infty}^{\infty} [p_i(\boldsymbol{\theta})]^{-1} \frac{\mathrm{d}p_{\mathbf{i}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \frac{\mathrm{d}p_{\mathbf{i}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^T}, \\
V(\boldsymbol{\theta}) &= I(\boldsymbol{\theta}) + 2 \sum_{t=2}^{p} \sum_{\mathbf{i}_1, \mathbf{i}_2} P(Y_1 \in A_{\mathbf{i}_1}, Y_t \in A_{\mathbf{i}_2}) p_{\mathbf{i}_1}^{-1} p_{\mathbf{i}_2}^{-1} \frac{\mathrm{d}p_{\mathbf{i}_1}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \frac{\mathrm{d}p_{\mathbf{i}_2}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^T}, \\
G(\boldsymbol{\theta}) &= I^{-1}(\boldsymbol{\theta}) V(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}).
\end{aligned}$$

Pak

$$\sqrt{n} \begin{pmatrix} \hat{c} - c \\ \hat{\boldsymbol{\phi}} - \boldsymbol{\phi} \\ \hat{\sigma^2} - \sigma^2 \end{pmatrix} \to^d N(0, G(\boldsymbol{\theta})),$$

takže odhad $(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma^2})$ *má asymptoticky mnohorozměrné normální rozdělení.*

Důkaz: Nechť $L_j(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma^2})$ je logaritmická věrohodnost založená na j-tém podvýběru $(\tilde{X}_j, ..., \tilde{X}_{p+j})$, $(\tilde{X}_{(p+1)+j}, ..., \tilde{X}_{(p+1)+p+j})$, ..., $(\tilde{X}_{(m-1)(p+1)+j}, ..., \tilde{X}_{(m-1)(p+1)+p+j})$, $j = 1, ..., p+1$, a

$$\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} = \left(\frac{\partial L_j(\boldsymbol{\theta})}{\partial \theta_1}, ..., \frac{\partial L_j(\boldsymbol{\theta})}{\partial \theta_{p+2}} \right)^T.$$

Po rozvinutí $\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$ v bodě $\boldsymbol{\theta}$ pomocí Taylorova vzorce dostáváme

$$0 = \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\bigg|_{\hat{\theta_j}} = \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} + (\boldsymbol{\theta_j} - \boldsymbol{\theta}) \frac{\mathrm{d}^2 L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}^2 \bigg|_{\hat{\theta_j^*}},$$

kde $\boldsymbol{\theta}_j^*$ je bod ležící na spojnici $\boldsymbol{\theta}$ a $\hat{\boldsymbol{\theta}}_j$. První rovnost vychází z konstrukce maximálně věrohodného odhadu.

Po dosazení maximálně věrohodného odhadu a jeho vyjádření z předchozího vzorce dostaneme

$$\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta} = \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \left(-\frac{\mathrm{d}^2 L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}}_j^*} \right)^{-1}.$$

Po zkonstruování odhadu průměrováním dílčích odhadů dostaneme

$$\frac{1}{p+1}\sum_{j=1}^{p+1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}-\boldsymbol{\theta})=\frac{1}{p+1}\sum_{j=1}^{p+1}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}|_{\hat{\theta}_j^*}\right)^{-1}.$$

Položme $m=\frac{n-p}{p+1}$. Na levé straně po vynásobení výrazem $\sqrt{n-p}$ a po sečtení dostáváme přibližně $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$, pokud je p oproti n zanedbatelné. Dílčí odhady $\hat{\boldsymbol{\theta}}_j$ jsou konstruovány na základě $p+1$ nezávislých výběrů a mají stejné rozdělení jako $\hat{\boldsymbol{\theta}}$. Tedy

$$\sqrt{n-p}\frac{1}{p+1}\sum_{j=1}^{p+1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{j}}-\boldsymbol{\theta})\approx\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}),$$

kde $\approx$ značí přibližnost. Označme

$$R=\frac{1}{p+1}\sum_{j=1}^{p+1}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}|_{\hat{\theta}_j^*}\right)^{-1}.$$

Úpravami dostaneme

$$\begin{aligned}R\sqrt{n-p}=\sqrt{n-p}\frac{1}{p+1}\sum_{j=1}^{p+1}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\hat{\theta}_j^*}\right)^{-1}\\ =\sqrt{m(p+1)}\frac{1}{p+1}\sum_{j=1}^{p+1}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\hat{\theta}_j^*}\right)^{-1}\\ =\frac{m}{\sqrt{m}}\frac{1}{\sqrt{p+1}}\sum_{j=1}^{p+1}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\hat{\theta}_j^*}\right)^{-1}.\end{aligned}$$

Výraz $\frac{m}{\sqrt{m}}$ vložíme do sumy a dostaneme

$$\frac{1}{\sqrt{p+1}}\sum_{j=1}^{p+1}\frac{1}{\sqrt{m}}\frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\left(-\frac{1}{m}\frac{\mathrm{d}^2L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2}\Big|_{\hat{\theta}_j^*}\right)^{-1},$$

čímž jsme odvodili vztah

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx \frac{1}{\sqrt{p+1}} \sum_{j=1}^{p+1} \frac{1}{\sqrt{m}} \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \left(-\frac{1}{m} \frac{\mathrm{d}^2 L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^2} \Big|_{\hat{\theta}_j^*} \right)^{-1}.$$

Bai a kol. (2009) uvádějí, že pro konzistentní odhad $\boldsymbol{\theta}_j$ konverguje výraz $-(\frac{1}{m})\frac{\mathrm{d}^2 L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\theta}}^T|_{\theta_j^*}$ k Fisherově matici $I(\boldsymbol{\theta})$ na základě $\tilde{X}_1, ..., \tilde{X}_{p+1}$. Označme $\boldsymbol{Y}_j = (\tilde{X}_j, ..., \tilde{X}_{p+j}), \qquad j = 1, ..., n-p.$

Na pravé straně dostaneme

$$\frac{1}{\sqrt{p+1}} \sum_{j=1}^{p+1} \frac{1}{\sqrt{m}} \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} I(\boldsymbol{\theta})^{-1} = I(\boldsymbol{\theta})^{-1} \frac{1}{\sqrt{p+1}} \frac{1}{\sqrt{m}} \sum_{j=1}^{p+1} \frac{\mathrm{d}L_j(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}},$$

Jelikož $m = \frac{n-p}{p+1}$, tak

$$\frac{1}{\sqrt{p+1}} \frac{1}{\sqrt{m}} \sim \sqrt{n},$$

hlavně pro $p << n$.

Součet prováděný přes podvýběry lze upravit na součet přes jednotlivé $(p+1)$-tice v podvýběrech a přes rozdělení prostoru odpovídající jednotlivým $(p+1)$-ticím. Tím dostaneme z $\sum_{j=1}^{p+1}$ sumy $\sum_{j=1}^{n-p} \sum_{\boldsymbol{i}} I_{(\boldsymbol{Y}_j \in A_{\boldsymbol{i}})}$, kde $I_{(bmY_j \in A_{\boldsymbol{i}})}$ je indikátor toho, že daná $(p+1)$-tice je v prostoru $A_{\boldsymbol{i}}$.

Pak můžeme použít vyjádření

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \frac{\boldsymbol{I}^{-1}(\boldsymbol{\theta})}{\sqrt{n}} \sum_{j=1}^{n-p} \sum_{\boldsymbol{i}} I_{(bmY_j \in A_{\boldsymbol{i}})} p_{\boldsymbol{i}}(\boldsymbol{\theta})^{-1} \frac{\mathrm{d}p_{\boldsymbol{i}}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}.$$

Součty p-závislých sekvencí jsou normovány a můžeme použít centrální limitní větu (CLV). Z CLV plyne, že $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to^L N[\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}) V_p(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta})]$ □.

Guo a Li (2012) odvozují korigovaný MLE odhad pro MA model.

MA(1) model uvažují ve tvaru $X_t = \varepsilon_t + \phi\varepsilon_{t-1}$, kde ε_t jsou nezávislé stejně rozdělené z $N(0, \sigma^2)$, $|\theta| < 1$. V tomto zápisu je náhodná veličina X_t centrovaná, tj. má nulovou střední hodnotu.

Z necentrovaného modelu

$$X_t = c + \sum_{l=0}^{p} \varepsilon_{t-l} \phi_l,$$

uvedeného Baiem a kol. (2009) ho dostaneme úpravou, kdy od náhodné veličiny X_t odečteme její střední hodnotu c. Tuto metodu lze použít i pro MA model vyššího řádu než 1.

Centrování MA(1) modelu má vliv pouze na střední hodnotu X_t a přítomnost aditivní konstanty c. Guo a Li (2012) se u MA(1) modelu věnují jen odhadům ϕ a σ^2, které nezávisí na centrování.

Sdruženou hustotu $\boldsymbol{X} = \boldsymbol{x}$, kde $\boldsymbol{X} = (X_1, ..., X_n)$ uvádějí Guo a Li (2012) jako

$$f(\boldsymbol{x}, \phi, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{t=1}^{n} \left(\sum_{i=0}^{t-1} \phi^i x_{t-i} \right)^2 \right\}. \tag{3.2}$$

Pro $t \leq 0$ předpokládají, že $x_t = 0$.

Guo a Li (2012) dále uvádějí, že věrohodnostní funce $\boldsymbol{\theta} = (\phi, \sigma^2)$ na základě zaokrouhlených dat $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_n)$ je definovaná vzorcem $L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) = h^{-n} \int_{\tilde{x}_n - h/2}^{\tilde{x}_n + h/2} ... \int_{\tilde{x}_1 - h/2}^{\tilde{x}_1 + h/2} f(\boldsymbol{u}, \boldsymbol{\theta}) \, \mathrm{d}u_1 ... \, \mathrm{d}u_n$ s hustotou uvedenou ve vzorci 3.2.

Korigovaný MLE odvozují první iterací Newtonovy metody $(\hat{\phi}_A, \hat{\sigma}_A^2)^T = (\hat{\phi}_0, \hat{\sigma}_0^2)^T - \mathbf{A}^{-1}\boldsymbol{b}$, kde

$$\mathbf{A} = \left| \begin{array}{cc} \frac{\partial^2 \ln f}{\partial \phi^2} & \frac{\partial^2 \ln f}{\partial \phi \partial \sigma^2} \\ \frac{\partial^2 \ln f}{\partial \phi \partial \sigma^2} & \frac{\partial^2 \ln f}{\partial (\sigma^2)^2} \end{array} \right|_{(\phi, \sigma^2) = (\tilde{\phi}_0, \tilde{\sigma}_0^2)},$$

$$\boldsymbol{b}^T = \frac{h^2}{24} \left(\sum_{t=1}^{n} \frac{\partial}{\partial \phi} \frac{\frac{\partial^2 f(\tilde{\boldsymbol{x}}, \phi, \sigma^2)}{\partial \tilde{x}_t^2}}{f(\tilde{\boldsymbol{x}}, \phi, \sigma^2)}, \sum_{t=1}^{n} \frac{\partial}{\partial \sigma^2} \frac{\frac{\partial^2 f(\tilde{\boldsymbol{x}}, \phi, \sigma^2)}{\partial \tilde{x}_t^2}}{f(\tilde{\boldsymbol{x}}, \phi, \sigma^2)} \right)_{(\phi, \sigma^2) = (\hat{\phi}_0, \hat{\sigma}_0^2)},$$

a $(\hat{\phi}_0, \hat{\sigma}_0^2)$ je pseudo maximálně věrohodný odhad (PMLE) parametru (ϕ, σ^2). Pseudo maximálně věrohodným odhadem parametru θ nazývají Guo a Li (2012) odhad získaný řešením rovnice $\frac{\mathrm{d} \ln f(x, \phi)}{\mathrm{d}\phi} = 0$, při dosazení zaokrouhlených dat, jako by šlo o přesná data (vycházel z něho i Lindley (1950)).

Následně uvádíme věty 3.5, 3.6 , na které se budeme odkazovat v důkazu věty .

Věta 3.5 *Mějme posloupnosti a_n a b_n, pro něž platí*

$$\lim_{n \to \infty} a_n = a \in R$$
$$\lim_{n \to \infty} b_n = b \in R.$$

Potom platí $\lim_{n \to \infty}(a_n + b_n) = a + b$.

Důkaz: Pro n-tý člen platí $|(a_n + b_n) - (a + b)| = |(a_n - a) + (b_n - b)| \leq |a_n - a| + |b_n - b|$. Buď ε libovolné kladné číslo, položme $\delta := \frac{1}{2}\varepsilon$. Pak existuje n_1, že pro každé $n \geq n_1$ je $|a_n - a| < \delta$. Dále existuje n_2, že pro každé $n \geq n_2$ je $|b_n - b| < \delta$. Položme $n_0 := \max(n_1, n_2)$. Pak pro každé $n \geq n_0$ platí $|(a_n + b_n) - (a + b)| < \varepsilon$. $\square$.

Věta 3.6 *Mějme posloupnosti a_n a b_n, $n \in N$ že platí* $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = a$. *Nechť existuje n_0, že pro každé $n \geq n_0$ platí $a_n \leq c_n \leq b_n$. Pak existuje* $\lim_{n\to\infty} c_n$ *a platí* $\lim_{n\to\infty} c_n = a$.

Důkaz: Důkaz uvádí Jarník (1974), v jeho knize se jedná o větu č.61 .
Následující větu používají v důkazu Guo a Li (2012), ale uvádějí ji bez důkazu.

Věta 3.7 *Mějme $\theta \in R, |\theta| < 1$. Pak platí*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=2}^{n} \sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2} = \frac{1+3\theta^2}{(1-\theta^2)^3}.$$

Důkaz: Pomocí softwaru Wolfram Alpha zjistíme, že

$$\begin{aligned}
&\sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2} \\
&= \frac{(2(t-1)^2 + 3(t-1) + 1)\theta^{2(t-1)} + (-4(t-1)^2 - 2(t-1) + 3)\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{(t-1)(2(t-1)-1)\theta^{2(t-1)+4} - 3\theta^2 - 1}{(\theta^2-1)^3} \\
&= \frac{(2t^2-t)\theta^{2t-2} + (-4t^2+6t+1)\theta^{2t} + (2t^2-5t+3)\theta^{2t+2} - 3\theta^2 - 1}{(\theta^2-1)^3} \\
&= \frac{(2t^2-t)\theta^{2t-2} + (-4t^2+6t+1)\theta^{2t} + (2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3} + \frac{-3\theta^2-1}{(\theta^2-1)^3}.
\end{aligned}$$

Platnost první rovnosti můžeme dokázat indukcí podle t:
1) Pro $t = 2$ má levá strana hodnotu $\sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2} = (2-1)\theta^{2-2} = 1$. Pravá strana má hodnotu

$$\begin{aligned}
&\frac{(2(t-1)^2 + 3(t-1) + 1)\theta^{2(t-1)} + (-4(t-1)^2 - 2(t-1) + 3)\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{(t-1)(2(t-1)-1)\theta^{2(t-1)+4} - 3\theta^2 - 1}{(\theta^2-1)^3} \\
&= \frac{6\theta^2 - 3\theta^4 + \theta^6 - 3\theta^2 - 1}{(\theta^2-1)^3} \\
&= \frac{\theta^6 - 3\theta^4 + 3\theta^2 - 1}{(\theta^2-1)^3} = 1,
\end{aligned}$$

tedy vzorec platí pro $t = 2$.
2) Indukční krok z t k $t + 1$:

$$
\begin{aligned}
&\sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2} \\
=& \frac{(2(t-2)^2+3(t-2)+1)\theta^{2(t-2)} + (-4(t-2)^2-2(t-2)+3)\theta^{2(t-2)+2}}{(\theta^2-1)^3} \\
&+ \frac{(t-2)(2(t-2)-1)\theta^{2(t-2)+4} - 3\theta^2 - 1}{(\theta^2-1)^3} \\
&+ (t-1)(2(t-1)-1)\theta^{2(t-1)-2}\frac{(\theta^2-1)^3}{(\theta^2-1)^3} \\
=& \frac{(2(t-2)^2+3(t-2)+1)\theta^{2(t-1)-2} + (-4(t-2)^2-2(t-2)+3)\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&+ \frac{(t-2)(2(t-2)-1)\theta^{2(t-1)+2} - 3\theta^2 - 1}{(\theta^2-1)^3} + (t-1)(2(t-1)-1)\frac{\theta^{2(t-1)+4}}{(\theta^2-1)^3} \\
&+ (t-1)(2(t-1)-1)\frac{(-3\theta^{2(t-1)+2} + 3\theta^{2(t-1)} - \theta^{2(t-1)-2})}{(\theta^2-1)^3} \\
=& \frac{[(2(t-2)^2+3(t-2)+1) - (t-1)(2(t-1)-1)]\theta^{2(t-1)-2}}{(\theta^2-1)^3} \\
&+ \frac{[(-4(t-2)^2-2(t-2)+3) + 3(t-1)(2(t-1)-1)]\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&+ \frac{[(t-2)(2(t-2)-1) - 3(t-1)(2(t-1)-1)]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&+ \frac{-3\theta^2-1}{(\theta^2-1)^3} + (t-1)(2(t-1)-1)\frac{\theta^{2(t-1)+4}}{(\theta^2-1)^3} \\
=& \frac{[(2(t^2-4t+4)+3t-5) - (2t^2-5t+3)]\theta^{2(t-1)-2}}{(\theta^2-1)^3} \\
&+ \frac{[(-4(t^2-4t+4) - (2t-4)+3) + (3(2t^2-5t+3))]\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&+ \frac{[(2t^2-9t+10) - (6t^2-15t+9)]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&+ \frac{-3\theta^2-1}{(\theta^2-1)^3} + (t-1)(2(t-1)-1)\frac{\theta^{2(t-1)+4}}{(\theta^2-1)^3}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{[(2t^2-8t+8+3t-5)-(2t^2-5t+3)]\theta^{2(t-1)-2}}{(\theta^2-1)^3} \\
&\quad + \frac{[(-4t^2+16t-16-2t+4+3)+(6t^2-15t+9)]\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&\quad + \frac{[2t^2-9t+10-6t^2+15t-9]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{-3\theta^2-1}{(\theta^2-1)^3} + (t-1)(2(t-1)-1)\frac{\theta^{2(t-1)+4}}{(\theta^2-1)^3} \\
&= \frac{[2t^2-t]\theta^{2(t-1)}}{(\theta^2-1)^3} + \frac{[-4t^2+6t+1]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{-3\theta^2-1}{(\theta^2-1)^3} + (t-1)(2(t-1)-1)\frac{\theta^{2(t-1)+4}}{(\theta^2-1)^3} \\
&= \frac{[2(t-1)^2+4t-2-t]\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&\quad + \frac{[-4(t-1)^2-8t+4+6t+1]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{(t-1)(2(t-1)-1)\theta^{2(t-1)+4}-3\theta^2-1}{(\theta^2-1)^3} \\
&= \frac{[2(t-1)^2+3(t-1)+1]\theta^{2(t-1)}}{(\theta^2-1)^3} \\
&\quad + \frac{[-4(t-1)^2-2(t-1)+3]\theta^{2(t-1)+2}}{(\theta^2-1)^3} \\
&\quad + \frac{(t-1)(2(t-1)-1)\theta^{2(t-1)+4}-3\theta^2-1}{(\theta^2-1)^3}
\end{aligned}
$$

Výraz $\frac{-3\theta^2-1}{(\theta^2-1)^3} = \frac{3\theta^2+1}{(1-\theta^2)^3}$ ve vzorci

$$
= \frac{(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3} + \frac{-3\theta^2-1}{(\theta^2-1)^3}.
$$

nezávisí na t ani na n a proto ho lze přičíst k limitě sumy zbývající části výrazu.

Nyní budeme vyšetřovat výraz

$$
\lim_{n\to\infty} \frac{1}{n} \sum_{t=2}^{n} \frac{(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3}. \tag{3.3}
$$

Postupným vyjádřením dostáváme

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}\frac{(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3}$$

$$=\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}[(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}].$$

Omezení zdola: Sčítáme přes t od 2 do n, t nahradíme n (uvažujeme ze součtu jen členy pro nejvyšší hodnotu t), tedy

$$\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}[(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}]$$

$$>\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}[(2n^2-n)\theta^{2n-2}+(-4n^2+6n+1)\theta^{2n}+(2n^2-5n+3)\theta^{2n+2}].$$

Pro $n\to\infty$ je n^2 řádově větší než n nebo konstanta, proto platí

$$\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}[(2n^2-n)\theta^{2n-2}+(-4n^2+6n+1)\theta^{2n}+(2n^2-5n+3)\theta^{2n+2}]$$

$$=\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}[(2n^2)\theta^{2n-2}+(-4n^2)\theta^{2n}+(2n^2)\theta^{2n+2}]$$

$$=\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}[(2n)\theta^{2n-2}+(-4n)\theta^{2n}+(2n)\theta^{2n+2}]$$

$$=\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}[2n\theta^{2n-2}-4n\theta^{2n}+2n\theta^{2n+2}].$$

Výraz θ^{2n-2} konverguje s rostoucím n k nule (protože $|\theta|<1$), $2n$ konvergují s rostoucím n k nekonečnu. Jelikož θ^{2n-2} konverguje rychleji, je limita prvního sčítance rovna nule. Podle stejného principu konvergují k nule i ostatní členy. Součet je definován, můžeme použít větu 3.5 a výraz konverguje k nule.

Omezení shora: Výraz

$$\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}[(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}]$$

omezíme shora tak, že ve výrazu nahradíme t za n a výraz přenásobíme počtem sčítanců určeného sumou $(n-1)$, tedy

$$\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}[(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}]$$

$$<\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}(n-1)[(2n^2-n)\theta^{2n-2}+(-4n^2+6n+1)\theta^{2n}+(2n^2-5n+3)\theta^{2n+2}].$$

Pro n jdoucí k nekonečnu je výraz n^2 řádově vyšší než n nebo konstanta a $\frac{n-1}{n}$ konverguje k jedné. Dostáváme

$$\begin{aligned}
&\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}\frac{1}{n}(n-1)[(2n^2-n)\theta^{2n-2}+(-4n^2+6n+1)\theta^{2n}+(2n^2-5n+3)\theta^{2n+2}] \\
&=\frac{1}{(\theta^2-1)^3}\lim_{n\to\infty}[2n^2\theta^{2n-2}-4n^2\theta^{2n}+2n^2\theta^{2n+2}].
\end{aligned}$$

Stejně jako při omezení zdola, výraz n^2 konverguje pro rostoucí n k nekonečnu a výraz θ^{2n-2} pro $|\theta|<1$ jde k nule, ale θ^{2n-2} konverguje rychleji než n^2. Všechny tři sčítance konvergují k nule a podle věty 3.5 konveruje k nule i výsledná limita.

Protože jsme nalezli vyšší i nižší výraz než 3.3, který konverguje ke stejné hodnotě (0), podle věty 3.6 je i limita výrazu 3.3 nulová.

Platí tedy

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-1}l(2l-1)\theta^{2l-2} \\
&=\lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}\left[\frac{(2t^2-t)\theta^{2t-2}+(-4t^2+6t+1)\theta^{2t}+(2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3}\right] \\
&+\frac{-3\theta^2-1}{(\theta^2-1)^3} \\
&=\frac{3\theta^2+1}{(1-\theta^2)^3}.\square
\end{aligned}$$

Následující tři věty uvádějí a dokazují Guo a Li (2012).

Věta 3.8 *Nechť*

$$\begin{aligned}
e_t &= \sum_{i=0}^{t-1}\phi^i\tilde{x}_{t-i}, \\
\dot{e}_t &= \sum_{i=1}^{t-1}i\phi^{i-1}\tilde{x}_{t-i}, \\
\ddot{e}_t &= \sum_{i=2}^{t-1}i(i-1)\phi^{i-2}\tilde{x}_{t-i}
\end{aligned}$$

a

$$\gamma_0^* = \gamma_0 + \frac{h^2}{12}$$
$$\gamma_0 = \sigma^2(1+\phi^2)$$
$$\gamma_1 = -\phi\sigma^2.$$

Potom platí:

$$\begin{aligned} &1)\ \lim_{n\to\infty}\frac{1}{n}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t) = \frac{(1+3\phi^2)\gamma_0^* + 2\phi(3+\phi^2)\gamma_1}{(1-\phi^2)^3}, \\ &2)\ \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{i,j=0}^{n-t}\phi^{i+j}e_{t+i}\dot{e}_{t+j} = \frac{\phi(2+\phi^2)\gamma_0^* + (1+5\theta^2)\gamma_1}{(1-\phi^2)^4}, \end{aligned} \tag{3.4}$$

Důkaz:

$$\begin{aligned} \frac{1}{n}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t) = &\frac{1}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-1} l(2l-1)\phi^{2l-2}\tilde{x}_{t-l}^2 \\ &+ \frac{2}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-2} l(2l+1)\phi^{2l-1}\tilde{x}_{t-l}\tilde{x}_{t-l+1} + R_n, \end{aligned}$$

kde

$$R_n = \frac{1}{n}\sum_{k=2}^{n-1}\sum_{t=k}^{n}\sum_{l=1}^{t-k} a_{l,t}(\phi)\tilde{x}_{t-l}\tilde{x}_{t-l+k}$$

s

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=k}^{n}\sum_{l=1}^{t-k} a_{l,t}(\phi) < \infty \tag{3.5}$$

pro všechna $2 \le k \le n-1$.

Z věty 3.7 a z Čebyševovy věty plyne

$$\frac{1}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-1} l(2l-1)\phi^{2l-2}y_{t-l}^2 \to_{n\to\infty} \frac{1+3\phi^2}{(1-\phi^2)^3}\gamma_0^* \tag{3.6}$$

v pravděpodobnosti,

kde

$$\begin{aligned} \gamma_0^* = E(\tilde{X}_i^2) = E(X_i+U_i)^2 &= \gamma_0 + \frac{h^2}{12} \\ \gamma_0 = E(X_i^2) &= \sigma_0^2(1+\phi^2), \end{aligned}$$

a U_i jsou zaokrouhlovací chyby.

Podobně

$$\frac{2}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-2} l(2l+1)\phi^{2l-1}\tilde{x}_i\tilde{x}_{i+1} \to_{n\to\infty} \frac{2\phi(3+\phi^2)}{(1+\phi^2)^3}\gamma_1^* \tag{3.7}$$

v pravděpodobnosti,

kde

$$\gamma_1 = E(\tilde{X}_t\tilde{X}_{t+1}) = E(X_tX_{t+1}) = -\phi_0\sigma_0^2.$$

Když

$$E(\tilde{X}_t\tilde{X}_{t+k}) = E(X_t+U_t)(X_{t+k}+U_{t+k}) = 0, \qquad k \geq 2,$$

tak z 3.5 a Čebyševovy věty plyne

$$R_n \to_{n\to\infty} 0 \tag{3.8}$$

v pravděpodobnosti. Výsledek 1) ze vzorců 3.4 plyne z konvergujících výrazů (vzorců) 3.6, 3.7 a 3.8. Podobně dostaneme výsledek 2) ze vzorců 3.4. □

Věta 3.9 *Nechť*

$$\hat{\theta}_0 = (\hat{\phi}_0, \hat{\sigma}_0^2)$$

je PMLE (ϕ, σ^2) *pro MA(1) model ve tvaru* $X_t = \varepsilon_t - \phi\varepsilon_{t-1}$. *Pak*

$$\begin{aligned} &1)\ \lim_{n\to\infty}\frac{1}{n}\boldsymbol{A} = -\frac{1}{\hat{\sigma}_0^2}\operatorname{diag}\left(\frac{(1+3\hat{\phi}_0^2)(1+\hat{\phi}_0^2)\hat{\sigma}_0^2 + 2\hat{\phi}_0(3+\hat{\phi}_0^2)(-1+\hat{\phi}_0\hat{\sigma}_0^2)}{(1-\hat{\phi}_0^2)^3}, \frac{1}{2\hat{\sigma}_0^2}\right), \\ &2)\ \lim_{n\to\infty}\frac{1}{n}\boldsymbol{b}^T = \frac{h^2}{12\hat{\sigma}_0^4}\left(\frac{\hat{\phi}_0}{(1-\hat{\phi}_0^2)^2}, -\frac{1}{2(1-\hat{\phi}_0^2)}\right). \end{aligned} \tag{3.9}$$

Důkaz: 1) Z normální rovnice $\frac{\partial \ln f(\tilde{x},\phi)}{\partial\phi} = 0$ máme

$$\begin{aligned} -\frac{1}{\sigma^2}\sum_{t=2}^{n} e_t\dot{e}_t\Big|_{\hat{\theta}_0} &= 0, \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n} e_t^2\Big|_{\hat{\theta}_0} &= 0 \end{aligned}$$

Z toho plyne, že druhé parciální derivace $\ln f$ vzhledem k ϕ a σ^2 jsou

$$\frac{\partial^2 \ln f}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n} e_t^2 = -\frac{n}{2\hat{\sigma}_0^4},$$
$$\frac{\partial^2 \ln f}{\partial\phi\partial\sigma^2} = \frac{1}{\sigma^4}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t) = 0.$$

Navíc platí

$$\frac{\partial^2 \ln f}{\partial\phi^2} = -\frac{1}{\sigma^2}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t),$$

a tak z věty 3.8 a tvaru matice $\boldsymbol{A}$ plyne výsledek 1) ze vzorců 3.9.
2) Poznamenejme že:

$$\frac{\partial}{\partial\phi}\frac{\partial^2 \ln(f)}{\partial\tilde{x}_t^2} = -\frac{2\phi^{2n-2t+1}[\phi^{-2(n-t)+(n-t)\phi^2-(n-t-1)}]}{\sigma^2(1-\phi^2)^2},$$
$$\frac{\partial}{\partial\sigma^2}\frac{\partial^2 \ln(f)}{\partial\tilde{x}_t^2} = \frac{1}{\sigma^4}\sum_{i=0}^{n-t}\phi^{2i},$$
$$\frac{\partial}{\partial\phi}\left(\frac{\partial \ln(f)}{\partial\tilde{x}_t}\right)^2 = \frac{2}{\sigma^4}\sum_{i=0}^{n-t}\phi^i e_{t+i}\left[\sum_{j=1}^{n-t} j\phi^{j-1}e_{t+j} + \sum_{k=0}^{n-t}\phi^k\dot{e}_{t+k}\right],$$
$$\frac{\partial}{\partial\sigma^2}\left(\frac{\partial \ln(f)}{\partial\tilde{x}_t}\right)^2 = -\frac{2}{\sigma^6}\sum_{i=0}^{n-t}\phi^i e_{t+i}.$$

Z aproximace b_j v korigovaném MLE plyne

$$b_1 = \frac{h^2}{12\hat{\sigma}_0^4}\sum\left[(1-\hat{\sigma}_0^2)\left(\sum_{i=1}^{n-t} i\hat{\phi}_0^{2i-1}\right) + \left(\sum_{i=t}^{n}\hat{\phi}_0^{i-t}e_i\right)\left(\sum_{j=t}^{n}\hat{\phi}_0^{j-t}\dot{e}_j\right)\right]\Bigg|_{\theta=\hat{\theta}_0},$$
$$b_2 = -\frac{h^2}{24\hat{\sigma}_0^4}\frac{\hat{\phi}_0^{2+2n} - \hat{\phi}_0^2 + (1-\hat{\phi}_0^2)n}{(1-\hat{\phi}_0^2)^2}.$$

Z toho plynou limity $\frac{b_1}{n}$ a $\frac{b_2}{n}$. □
Z věty 3.9 a definice korigovaného MLE plyne následující věta:

Věta 3.10 *Korigovaný MLE modelu MA(1) založený na zaokrouhlených datech*

je

$$\hat{\phi}_A \approx \hat{\phi}_0 + \frac{\hat{\phi}_0 h^2}{12\hat{\sigma}_0^4}\left[\frac{1}{1-\hat{\phi}_0^2} - \frac{(1+\hat{\phi}_0^2)(1-\hat{\sigma}_0^2)}{n(1-\hat{\phi}_0^2)^2}\right],$$
$$\hat{\sigma}_A^2 \approx \hat{\sigma}_0^2 - \frac{h^2}{12}\left[\frac{1}{1-\hat{\phi}_0^2} - \frac{\hat{\phi}_0^2}{n(1-\hat{\phi}_0^2)^2}\right],$$

kde $\hat{\phi}_0, \hat{\sigma}_0^2$ *jsou PMLE a* h *je šířka zaokrouhlovacího intervalu.*

3.2 Model AR

Proces AR(p) lze vyjádřit zápisem

$$X_{t+1} = c + \sum_{l=1}^{p} \phi_l X_{t-l+1} + \varepsilon_{t+1},$$

kde ε_l jsou nezávislé stejně rozdělené chyby s normálním rozdělením $N(0, \sigma^2)$, $l = 1, ..., n$. Předpokládejme, že AR(p) proces je kauzální. Z předpokladu kauzality vyplývá, že $1 - \sum_{l=1}^{p} \phi_l z^l \neq 0$ pro $|z| \leq 1$. Pak jsou náhodné veličiny $(X_1, ..., X_n)$ normálně rozdělené s $N(\mu \mathbf{1}_n, \mathbf{\Sigma}_{n\times n})$, kde $\mathbf{1}_n = (1, 1, ..., 1)^T$ je vektor jedniček o n řádcích, $\mu = \frac{c}{1-\phi_1-...-\phi_p}$.

Nyní předpokládejme, že jsou k dispozici pouze data zaokrouhlená na celá čísla. Ta označíme $\tilde{X}_1, ..., \tilde{X}_n$, kde $\tilde{X}_i$ je zaokrouhlená hodnota X_i na celé číslo. Bai a kol. (2009) uvádí, že pokud jsou ignorovány zaokrouhlovací chyby, obvyklé odhady získané z řešení Yule-Walkerových rovnic nejsou konzistentní. Konzistenci odhadů získaných řešením Yule-Walkerových rovnic z nezaokrouhlených dat dokazují Brockwell a Davis (1991), v jejich publkaci se jedná o větu 8.1.1.

Potřebujeme najít odhady parametrů $\boldsymbol{\phi} = (c, \phi_1, .., \phi_p)^T$ a σ^2. Jednou z možností, jak odhad zkonstruovat, je aproximace metody maximální věrohodnosti.

Bai a kol. (2009) nejprve uvádějí odvození pro $p = 1$ (model AR(1)), nejprve se tedy budeme zabývat speciálně tímto modelem.

3.2.1 AR(1)

Pro AR(1) model máme vztah $X_{t+1} = c+\phi_1 X_t+\varepsilon_t$, kde $|\phi_1| < 1$ a ε_t jsou nezávislé stejně rozdělené chyby s $N(0, \sigma^2)$ pro $t = 1, ..., n-1$. Jelikož v AR(1) modelu máme pouze ϕ_1, budeme ho dále označovat pouze ϕ. Pak vektor $(X_1, ..., X_n)^T$ má normální rozdělení $N(\mu \mathbf{1_n}, \mathbf{\Sigma_{n\times n}})$, kde $\mu = \frac{c}{1-\phi}$ a

$$\mathbf{\Sigma}_{n\times n} = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & ... & \phi^{n-1} \\ \phi & 1 & \phi & ... & \phi^{n-2} \\ \phi^2 & \phi & 1 & ... & \phi^{n-3} \\ ... & ... & ... & ... & ... \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & ... & 1 \end{pmatrix}.$$

Konkrétně máme

$$\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi \\ \phi & 1 \end{pmatrix}\right),$$

kde ϕ je korelační koeficient mezi X_t a X_{t+1}.

Protože $|\phi| < 1$ a chceme, aby $|\phi|^k$ bylo hodně malé, můžeme si za tímto účelem zvolit vhodně velké k. Pokud je k dostatečně velké, (X_i, X_{i+1}) a (X_{i+k}, X_{i+k+1}) jsou přibližně nezávislé a tedy i dvojice zaokrouhlených dat $(\tilde{X}_i, \tilde{X}_{i+1})$ a $(\tilde{X}_{i+k}, \tilde{X}_{i+k+1})$ jsou přibližně nezávislé.

Na základě výše uvedeného argumentu Bai a kol. (2009) odvozují postup konstrukce aproximace maximálně věrohodného odhadu pro parametry c, ϕ a σ^2. Předpokládejme, že k je velké přirozené číslo. Rozdělme zaokrouhlená data na k podskupin o velikosti $m = \frac{n-1}{k}$. (Překrývání skupin je povoleno.)

Rozdělení na skupiny je následující:

1) $(\tilde{X}_1, \tilde{X}_2)$, $(\tilde{X}_{k+1}, \tilde{X}_{k+2})$,...,$(\tilde{X}_{(m-1)k+1}, \tilde{X}_{(m-1)k+2})$

2) $(\tilde{X}_2, \tilde{X}_3)$, $(\tilde{X}_{k+2}, \tilde{X}_{k+3})$,...,$(\tilde{X}_{(m-1)k+2}, \tilde{X}_{(m-1)k+3})$

... .

k) $(\tilde{X}_k, \tilde{X}_{k+1})$, $(\tilde{X}_{2k}, \tilde{X}_{2k+1})$,...,$(\tilde{X}_{mk}, \tilde{X}_{mk+1})$

Bai a kol. (2009) uvádějí následující konstrukci odhadu na základě zaokrouhlených dat $\tilde{X}_1, ..., \tilde{X}_n$:

1) Uvažujeme $(\tilde{X}_1, \tilde{X}_2)$,$(\tilde{X}_{k+1}, \tilde{X}_{k+2})$,...,$(\tilde{X}_{(m-1)k+1}, \tilde{X}_{(m-1)k+2})$ jako výběr nezávislých stejně rozdělených dvoudimenzionálních vektorů. Označme $n_{ij}^{(l)}$ počet výskytů dvojic $(i, j), i < j$ v l-té skupině $(l = 1, ..., k)$. Pomocí metody nelineárního programování maximalizujeme aproximaci logaritmické věrohodnosti $\sum_{ij} n_{ij}^{(1)} \ln p_{ij}$, kde $\sum_{ij} p_{ij} = 1$. Metodu nelineárního programování uvádí např. Lachout (2008).

Pak zkonstruujeme aproximaci maximálně věrohodného odhadu parametrů (c, ϕ, σ^2) a označíme je $(\hat{c}_1, \hat{\phi}_1, \hat{\sigma}_1^2)$, kde $n_{ij}^{(1)}$ jsou četnosti výskytů dvojic (i, j) v prvním podvýběru.

2) Podobně zkonstruujeme aproximaci maximálně věrohodného odhadu na základě 2. až m-tého podvýběru a odhady parametrů označíme $(\hat{c}_j, \hat{\phi}_j, \hat{\sigma}_j^2)), j = 2, ..., m$.

3) Konečnou aproximaci maximálně věrohodného odhadu získáme zprůměrováním k dílčích odhadů:

$$\hat{c} = \sum_{i=1}^{k} \frac{\hat{c}_i}{k},$$

$$\hat{\phi} = \sum_{i=1}^{k} \frac{\hat{\phi}_i}{k},$$

$$\hat{\sigma}^2 = \sum_{i=1}^{k} \frac{\hat{\sigma}_i^2}{k}.$$

Věta 3.11 *Pokud zaokrouhlíme veličiny generované AR(1) modelem, budou se chovat jako nezaokrouhlené veličiny v modelu ARMA(1,1).*

Při dokazování využijeme odvození při oddělování signálu a šumu, které uvádí Anděl (1976).

Mějme model s autokovarianční funkcí $Ca^{|t|}, C > 0, a \neq 0, a \in (-1,1)$ a nekorelovaný šum s nulovou střední hodnotou a kladným rozptylem D. Signál s autokovarianční funkcí $Ca^{|t|}$ má spektrální hustotu $f_x(\lambda) = \frac{C(1-a^2)}{2\pi}\frac{1}{|e^{i\lambda}-a|^2}$ a šum Y_t má spektrální hustotu $f_y(\lambda) = \frac{D}{2\pi}$.

Pokud je šum se signálem nekorelovaný, má jejich součet $f_z(\lambda) = f_x(\lambda)+f_y(\lambda)$ spektrální hustotu $f_z(\lambda) = B\frac{|e^{i\lambda}-b|^2}{|e^{i\lambda}-a|^2}$, kde $B = \frac{C}{2\pi}$ a $b \neq 0, b \in (-1,1), b \neq a$.

Anděl (1976) uvádí ARMA posloupnost v nobecnějším tvaru $\sum_{i=0}^{p} a_i X_{t-i} = \sum_{j=0}^{q} b_j \varepsilon_{t-j}$. a spektrální hustotu ve tvaru $f(\lambda) = \frac{\sigma^2}{2\pi}\frac{|b_0+b_1e^{-i\lambda}+\ldots+b_qe^{-im\lambda}|^2}{|a_0+a_1e^{-i\lambda}+\ldots+a_pe^{-in\lambda}|^2}$, zatímco jinde bývá uváděn navíc s podmínkou $a_0 = b_0$.

Model AR(1) je speciálním případem modelu s autokovarianční funkcí $Ca^{|t|}$, pro $C = \frac{\sigma^2}{1-a^2}, \sigma^2 > 0$.

Anděl pokládá $A = \frac{C(1-a^2)}{2\pi}$ a rozkládá $f_z(\lambda)$ na

$$\begin{aligned} f_z(\lambda) &= B\frac{|e^{i\lambda}-b|^2}{|e^{i\lambda}-a|^2} \\ &= B\left[\frac{b}{a} + \frac{(a-b)(1-ab)}{a}\frac{1}{(e^{i\lambda}-a)(e^{-i\lambda}-a)}\right]. \end{aligned}$$

Uvedená spektrální hustota $f_z(\lambda) = B\frac{|e^{i\lambda}-b|^2}{|e^{i\lambda}-a|^2}$ je pro $C = \frac{\sigma^2}{1-a^2}$ rovna

$$f_z(\lambda) = \frac{\sigma^2}{2\pi(1-a^2)}\frac{|e^{i\lambda}-b|^2}{|e^{i\lambda}-a|^2}.$$

Nyní ukážeme, že to odpovídá spektrální hustotě modelu ARMA(1,1) s rozptylem $\frac{\sigma^2}{1-a^2}$, protože

$$\begin{aligned} \frac{|e^{i\lambda}-b|^2}{|e^{i\lambda}-a|^2} &= \frac{|\frac{e^{-i\lambda}}{e^{-i\lambda}}(e^{i\lambda}-b)|^2}{|\frac{e^{-i\lambda}}{e^{-i\lambda}}(e^{i\lambda}-a)|^2} = \frac{|(\frac{1}{e^{-i\lambda}})(1-e^{-i\lambda}b)|^2}{|(\frac{1}{e^{-i\lambda}})(1-e^{-i\lambda}a)|^2} \\ &= \frac{|(\frac{1}{e^{-i\lambda}})|^2|(1-e^{-i\lambda}b)|^2}{|(\frac{1}{e^{-i\lambda}})|^2|(1-e^{-i\lambda}a)|^2} = \frac{|(1-e^{-i\lambda}b)|^2}{|(1-e^{-i\lambda}a)|^2}. \end{aligned}$$

Hodnoty parametrů $a_0 = 1, b_0 = 1, a_1 = -a, b_1 = -b$ vyhovují i obvykle uváděnému tvaru modelu ARMA(1,1), kde $a_0 = b_0$.

Anděl (1976) uvádí, že ke splnění rovnosti $f_z(\lambda) = f_x(\lambda) + f_y(\lambda)$ musí platit

$$\begin{aligned} \frac{Bb}{a} &= \frac{D}{2\pi} \\ \frac{B(a-b)(1-ab)}{a} &= A. \end{aligned}$$

Dále též uvádí úpravu na kvadratickou rovnici v b. Vydělením rovnic výrazem $2\pi A$ a úpravou dostaneme $b = \frac{D(a-b)(1-ab)}{2\pi A}$. Výraz upravíme a vyřešíme kvadratickou rovnici pro b.

Vyjádření kvadratické rovnice vzhledem k b:

$$\begin{aligned} b &= \frac{D(a-b)(1-ab)}{2\pi A} \\ b2\pi A &= D(a-b)(1-ab) \\ b2\pi A &= D(a-b-a^2b+ab^2) \\ b2\pi A &= Da - bD - Da^2b + Dab^2 \\ 0 &= Da - bD - Da^2b + Dab^2 - b2\pi A \end{aligned}$$

Nahradíme vyjádření za A ($A = \frac{C(1-a^2)}{2\pi}$).

$$\begin{aligned} 0 &= b^2Da - b(D + 2\pi\frac{C(1-a^2)}{2\pi} + Da^2) + Da \\ 0 &= b^2Da - b(D + C(1-a^2) + Da^2) + Da \end{aligned}$$

Dále odvozujeme některé vlastnosti řešení výše uvedené kvadratické rovnice. Vyjáříme diskriminant ke kvadratické rovnici. Označme ho R.

$$\begin{aligned} R &= (D(1+a^2) + C(1-a^2))^2 - 4D^2a^2 \\ &= D^2(1+2a^2+a^4) + 2D(1+a^2)C(1-a^2) + C^2(1-a^2)^2 - 4D^2a^2 \\ &= D^2(1-a^2)^2 + 2D(1+a^2)C(1-a^2) + C^2(1-a^2)^2 \end{aligned}$$

Diskriminant doplníme na čtverec:

$$\begin{aligned} R &= D^2(1-a^2)^2 + 2D(1+a^2)C(1-a^2) + C^2(1-a^2)^2 \\ &= (D(1-a^2) + C(1-a^2))^2 - 2CD(1-a^2)^2 + 2D(1+a^2)C(1-a^2) \\ &= (D(1-a^2) + C(1-a^2))^2 + 4CD(1-a^2)a^2 \end{aligned}$$

Druhá mocnina (první člen výrazu R) je kladná a z omezení na a a kladnosti C a D je kladný i druhý člen R, tedy diskriminant je kladný.

$$\begin{aligned} b &= \frac{(D + C(1-a^2) + Da^2) \pm \sqrt{R}}{2Da} \\ b_1 &= \frac{(D + C(1-a^2) + Da^2) + \sqrt{R}}{2Da} \\ b_2 &= \frac{(D + C(1-a^2) + Da^2) - \sqrt{R}}{2Da} \end{aligned}$$

Nyní ukážeme, že součin kořenů je 1:

$$
\begin{aligned}
b_1 * b_2 &= \frac{(C(1-a^2)+D(1+a^2))+\sqrt{R}}{2Da} * \frac{(D+C(1-a^2)+Da^2)-\sqrt{R}}{2Da} \\
&= \frac{(C(1-a^2)+D(1+a^2))^2-R}{4D^2a^2} \\
&= \frac{(C(1-a^2)+D(1+a^2))^2-(D(1-a^2)+C(1-a^2))^2-4CD(1-a^2)a^2}{4D^2a^2} \\
&= \frac{C^2(1-a^2)^2+2CD(1-a^2)(1+a^2)+D^2(1+a^2)^2}{4D^2a^2} \\
&+ \frac{-[D^2(1-a^2)^2+2CD(1-a^2)^2+C^2(1-a^2)^2]-4CD(a^2-a^4)}{4D^2a^2} \\
&= \frac{2CD(1-a^2)(1+a^2)+D^2(1+a^2)^2}{4D^2a^2} \\
&+ \frac{-D^2(1-a^2)^2-2CD(1-a^2)^2-4CD(a^2-a^4)}{4D^2a^2} \\
&= \frac{2CD(1-a^4)+D^2(1+2a^2+a^4)}{4D^2a^2} \\
&+ \frac{-D^2(1-2a^2+a^4)-2CD(1-2a^2+a^4)-4CD(a^2-a^4)}{4D^2a^2} \\
&= \frac{2CD(1-a^4)+D^2(2a^2)}{4D^2a^2} \\
&+ \frac{-D^2(-2a^2)-2CD(1-2a^2+a^4)-4CD(a^2-a^4)}{4D^2a^2} \\
&= \frac{CD(2(1-a^4)-2(1-2a^2+a^4)-4(a^2-a^4))}{4D^2a^2} \\
&+ \frac{-D^2(-2a^2)+D^2(2a^2)}{4D^2a^2} \\
&= \frac{2CD(-a^4+2a^2-a^4-2a^2+2a^4)}{4D^2a^2} + \frac{4D^2a^2}{4D^2a^2} \\
&= \frac{4D^2a^2}{4D^2a^2} = 1.
\end{aligned}
$$

Jelikož součin kořenů je 1 a kořeny jsou různé reálné, mají oba stejné znaménko a jeden z nich musí být v absolutní hodnotě menší než 1. Tím je potvrzeno, že jeden z kořenů b_1, b_2 splňuje podmínky pro parametr b pro výše uvedený tvar spektrální hustoty modelu s šumem.

Kořeny b vypočítáme tak, že výraz

$$(D+C(1-a^2)+Da^2) \pm \sqrt{(D(1-a^2)+C(1-a^2))^2+4CD(1-a^2)a^2})$$

dělíme $2Da$, tedy

$$b_1 = \frac{(C(1-a^2)+D(1+a^2)) + \sqrt{(D(1-a^2)+C(1-a^2))^2 + 4CD(1-a^2)a^2}}{2Da},$$
$$b_2 = \frac{(C(1-a^2)+D(1+a^2)) - \sqrt{(D(1-a^2)+C(1-a^2))^2 + 4CD(1-a^2)a^2}}{2Da}.$$

Rozptyl šumu D je kladný, takže b má stejné znaménko jako a, pokud je kladný výraz v čitateli. Vzhledem ke stejnému znaménku kořenů stačí ukázat, že je kladný čitatel pouze pro jeden kořen.

Vezměme čitatel kořene b_1. První část výrazu, $C(1-a^2)+D(1+a^2)$ je kladná, protože C i D jsou kladné a $|a| < 1$, takže obě jsou přenásobeny kladným číslem ($1 - a^2$ nemůže s takto omezeným a nabýt záporné hodnoty). Diskriminant je kladný, jak jsme již ukázali. Jeho odmocnina je také kladné číslo a součet samých kladných čísel je kladný. Kořeny tedy mají stejné znaménko jako a.

Ještě zbývá vyšetřit, zda jsou zaokrouhlovací chyby v AR(1) modelu nekorelované s hodnotami veličiny X. Aby byly nekorelované, musí platit $E(XU) = 0$, neboť střední hodnota zaokrouhlovacích chyb je nulová.

Předpokládáme, že veličiny X mají normální rozdělení. Jelikož kovarance se s přičtením konstanty k veličině nezmění, můžeme bez újmy na obecnosti uvažovat normální rozdělení s nulovou střední hodnotou.

Každé hodnotě veličiny je (v závislosti na hrubosti zaokrouhlení) jednoznačně přiřazena její zaokrouhlená hodnota. Toho využijeme při odvození $E(XU)$ z klasického vzorce pro střední hodnotu spojité veličiny $EY = \int_{-\infty}^{\infty} yf(y)\,\mathrm{d}y$. Protože zaokrouhlovací chyby nejsou spojité na celém R, ale na jejích podintervalech, budeme sčítat přes tyto intervaly. Na jednotlivých intervalech vyjádříme lineární vztah mezi hodnotou veličiny a zaokrouhlovací chybou.

Získáme tak konečný vzorec

$$\sum_{n=-\infty}^{\infty} \int_{(n-1/2)h}^{(n+1/2)h} f(x)x((nh-x)/h)\,\mathrm{d}x. \tag{3.10}$$

Pro šířku intervalu $h = 1$ se vzorec zjednoduší na

$$\sum_{n=-\infty}^{\infty} \int_{(n-1/2)}^{(n+1/2)} f(x)x(n-x)\,\mathrm{d}x,$$

tento výraz však nejsme schopni vypočítat.

V programu R se pokusíme ilustrovat, jaký by mohl být výsledek vzorce 3.10. Funkce v programu R budeme značit `tímto fontem`. Pomocí funkce `integrate` v programu R zjistíme integrály přes jednotlivé intervaly $[(n-1/2)h; (n+1/2)h)$. Jelikož v R nemůžeme sčítat nekonečně mnoho sčítanců, ukážeme si chování na několika oblastech centrovaných kolem nuly. Výsledky jsou uvedeny v tabulce 3.1.

Pomocí výpočtu v programu R se nám sice nepovedlo dokázat, že normálně rozdělená veličina je nekorelovaná se zaokrouhlovacími chybami, nicméně na

Tabulka 3.1: Porovnání vypočítaných korelací veličiny X a zaokrouhlení U

σ^2 a h	(1;1)	(1;0,1)	(5;1)	(5;0,1)
-1 až 1	-0,03273735	-0,0009866883	-0,01910656	-0,0001993854
-10 až 10	-5,350576e-09	-0,004024211	-0,01537969	-0,001365873
-100 až 100	-5,350576e-09	4,294697e-17	6,092372e-18	-0,001772782
-1000 až 1000	-5,350576e-09	4,294697e-17	6,092372e-18	7,287922e-16
-10000 až 10000	-5,350576e-09	4,294697e-17	6,092372e-18	7,287922e-16

V této tabulce porovnáváme hodnoty $\sum_{n=-a}^{b} \int_{(n-1/2)}^{(n+1/2)} f(x)x(n-x)\,\mathrm{d}x$, protože v programu R nemůžeme sečíst nekonečný počet hodnot. Meze a a b jsou vypsány v prvním sloupci (platí pro celý řádek), hodnoty rozptylu a šířka zaokrouhlovacího intervalu jsou uvedeny v prvním řádku (platí pro celý sloupec). Přestože výsledné hodnoty nejsou nulové, jsou blízké nule. Trend s rozšiřující se oblastí výpočtu (nižší řádky v tabulce) nenaznačuje, že by hodnota korelace veličiny a zaokrouhlení rostla.

základě získaných výsledků bychom mohli očekávat, že jejich korelace je velmi malá a nedopustíme se velké chyby, když se k šumu AR(1) modelu budeme chovat, jako by s veličinou X nekorelovaný byl. □

Chování AR(1) modelu na nezaokrouhlených a zaokrouhlených datech ukážeme na simulaci.

Nasimulovali jsme pro model AR(1) hodnoty $X_i, i = 1, ..., 1000$ s ε_i, $i = 1, ..., 1000$ z N(0,1), $\phi_1 = 0,5$, $c = 0$ s X_0 náhodně vybraného z N(0,1) a výsledky jsme zaokrouhlili na celá čísla. Zaokrouhlená data jsme vyšetřovali pomocí funkcí v softwaru R, nejen mezi ARMA modely, ale i mezi užší skupinou AR modelů nebo širší skupinou ARIMA modelů, v některých případech přímo konkrétní model (AR(1) nebo ARMA(1,1)). Nejprve jsou uvedeny výsledky u vyšetřování modelů AR a ARMA jednotlivými funkcemi. Následuje shrnutí a komentář výsledků, porovnání výsledků v modelech AR a ARMA a porovnání věrohodnostních metod a metody momentů.

V ARMA modelu budeme značit členy pro AR pomocí δ a člen pro MA pomocí ϕ. Jelikož v tomto případě vyšetřujeme model AR(1) a na zaokrouhlených datech očekáváme model ARMA(1,1), budeme dále členy uvádět bez dolního indexu 1.

Nejprve budeme testovat hypotézu, že kovariance dat z N(0,1) s jejich zaokrouhlenými hodnotami je nulová na hladině 0,05. Testování t-testem (`t.test`) neprokázalo závislost zaokrouhlovacích chyb s daty, výsledná p-hodnota byla přibližně 0,49 .

Přikročíme tedy k vlasnímu vyšetřování modelů ARMA a AR. V programu R jsme testovali hodnoty koeficientů v modelu ARMA(1,1) pomocí funkce `arma`, která odhad provádí na základě podmíněných nejmenších čtverců.

Pro zaokrouhlená data měl odhad pro δ hodnotu $0,509896$, pro ϕ hodnotu $-0,041563$ a pro konstantu b hodnotu $0,002915$. Podmíněný součet čtverců byl 1116,38 a hodnota Akaikeho kritéria byla 2955,97. Při vyšetření nezaokrouh-

lených dat jsme získali výsledky $\delta = 0,512525$, $\phi = 0,006958$ a $b = -0,004834$. Podmíněný součet čtverců byl 994,77 a hodnota Akaikeho kritéria byla 2840,64.

Při hledání vhodného modelu pomocí Akaikeho kritéria pro model ARMA(p,q) pomocí funkce `arma` s omezením pro hodnoty $p, q \leq 5$, vyšla jak pro nezaokrouhlená, tak pro zaokrouhlená data nejnižší hodnota AIC pro model AR(1), i když odhadnutá hodnota parametru ϕ na zaokrouhlených datech je řádově větší. Pro nezaokrouhlená data mělo Akaikeho kritérium hodnotu 2838,653, pro zaokrouhlená data mělo hodnotu 2954,33.

Při testování modelu ARMA(1,1) pomocí funkce `arima`, která odhad provádí na základě maximální věrohodnosti, měl pro zaokrouhlená data odhad pro δ hodnotu $0,5093$, pro ϕ hodnotu $-0,0414$ a pro konstantu b hodnotu $0,0061$. Při vyšetření nezaokrouhlených dat jsme získali výsledky $\delta = 0,5116$, $\phi = 0,0076$ a $b = -0,0112$. I v tomto případě je odhadnutá hodnota parametru ϕ na zaokrouhlených datech řádově větší.

Při hledání vhodného modelu pomocí Akaikeho kritéria vychází v tomto případě pro model ARMA(p,q) s hodnotami $p, q \leq 5$ nejnižší hodnota (2833,34) pro model ARMA(4,5) pro nezaokrouhlená data. Pro zaokrouhlená data vychází v tomto ohledu jako nejlepší model ARMA(4,2), s hodnotou Akaikeho kritéria 2954,497.

Při vyšetřování statistik modelu ARMA(1,1) na základě věrohodnostní metody McLeoda a Zhanga (2007) funkcí `FitARMA` získáváme pro nezaokrouhlená data logaritmickou věrohodnost 2,4 a hodnotu Akaikeho kritéria 1,2. Pro zaokrouhlená data vychází logaritmická věrohodnost $-55,18$ a hodnota Akaikeho kritéria 116,4. Pro zaokrouhlená data byl odhad δ roven 0,5093646, odhad ϕ byl 0,04148934 a odhad střední hodnoty řady byl 0,007. Pro nezaokrouhlená data byl odhad δ roven 0,5115869, odhad ϕ byl $-0,007589097$ a odhad střední hodnoty řady byl $-0,009421167$.

Při hledání nejvhodnějšího modelu mezi ARMA modely pro $p, q \leq 5$ dostáváme v obou případech model AR(1), přestože některé z modelů nelze danou metodou otestovat. Přitom některé modely nejde otestovat pouze na nezaokrouhlených datech - na zokrouhlených otestovat jdou. Pro nezaokrouhlená data je hodnota Akaikeho kritéria přibližně $-0,789$, pro zaokrouhlená data má Akaikeho kritérium hodnotu přiblžně 114,710.

Při vyšetřování statistik AR(1) modelů věrohodnostní metodou McLeoda a Zhanga(2007) dostáváme pro nezaokrouhlená data logaritmickou věrohodnost 2,39 a Akaikeho kritérium $-0,8$, pro zaokrouhlená data dostáváme logaritmickou věrohodnost -55,36 a Akaikeho kritérium 114,7.

Odhad parametru δ vychází 0,4773745 pro zaokrouhlená data a 0,5171466 pro data nezaokrouhlená. Odhad střední hodnoty řady je 0,007 pro zaokrouhlená data a $-0,009421167$ pro data nezaokrouhlená.

Při vyšetřování pouze mezi AR modely do řádu 5 pomocí Yule-Walkerovy metody funkcí `ar` dostaneme v obou případech model AR(1). Pro nezaokrouhlená data nám vyjde s hodnotou koeficientu δ rovna 0,517 a pro zaokrouhlená data vyjde hodnota koeficientu 0,4775.

Při vyšetřování metodou maximální věrohodnosti do řádu 5 funkcí `ar` vyjde v obou případech také jako nejlepší model AR(1). Pro nezaokrouhlená data

vychází koeficient δ 0,5171, pro zaokrouhlená $0,4774$.

Při odhadu koeficientu ϕ AR(1) modelu při známé střední hodnotě pomocí funkce `AR1Est` vyšlo 0,5171785 pro nezaokrouhlená a 0,4773924 pro zaokrouhlená data.

Při vyšetřování statistik modelu AR(1) pomocí metody maximální věrohodnosti McLeoda a Zhanga (2006) funkcí `FitAR` získáme pro nezaokrouhlená data logaritmickou věrohodnost 2,394 a AIC $-0,8$, pro zaokrouhlená data logaritmickou věrohodnost $-55,355$ a AIC 114,7. Odhad parametru vychází 0,4773746 pro zaokrouhlená data a 0,5171466 pro nezaokrouhlená data.

Při odhadu koeficientů v modelu ARMA(1,1) vychází odhad koeficientu ϕ u nezaokrouhlených dat řádově vyšší. Přesto vychází na zaokrouhlených i nezaokrouhlených datech pro modely ARMA(p,q) pro $p, q \leq 5$ nejnižší Akaikeho kritérium pro model AR(1). (Kromě vyšetřování funkcí `arima`, která jako nejlepší vyhodnotila modely ARMA(4, 2) a ARMA(4, 5).)

Při vyšetřování mezi AR modely je model na zaokrouhlených i nazeokrouhlených datech správně identifikován jako AR(1). Hodnota Akaikeho kritéria byla pro nezaokrouhlená data menší, v některých případech dokonce i řádově. Věrohodnost byla naopak větší na nezaokrouhlených datech a menší na zaokrouhlených. Odhad parametru δ měl v ARMA(1,1) i AR(1) modelech menší odchylku od skutečné hodnoty na zaokrouhlených datech než na nezaokrouhlených.

Experimentálně se nám nepodařilo prokázat, že zaokrouhlená data generovaná modelem AR(1) se chovají jako data modelu ARMA(1,1).

3.2.2 Obecný AR model

Uvažujme AR(p) model $X_t = c + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \varepsilon_t$, kde ε_t jsou nezávislé chyby se stejným normálním rozdělením $N(0, \sigma^2)$ pro $t = p+1, ..., n$. Nechť $\boldsymbol{\theta} = (c, \phi_1, ..., \phi_p, \sigma^2)^T = (c, \boldsymbol{\phi}, \sigma^2)^T$. Napozorována jsou pouze zaokrouhlená data $\tilde{X}_1, ..., \tilde{X}_n$. Budeme konstruovat odhady $(\hat{c}, \hat{\boldsymbol{\phi}}, \hat{\sigma}^2)$. Nechť $m = \frac{n-p}{k}$. Výběr po vzoru Baie a kol. (2009) rozdělíme na k podvýběrů:

1) $\tilde{X}_1, ..., \tilde{X}_{p+1},\ \tilde{X}_{k+1}, ..., \tilde{X}_{k+p+1},\ ..., \tilde{X}_{(m-1)k+1}, ..., \tilde{X}_{(m-1)k+p+1}$

2) $\tilde{X}_2, ..., \tilde{X}_{p+2},\ \tilde{X}_{k+2}, ..., \tilde{X}_{k+p+2},\ ..., \tilde{X}_{(m-1)k+2}, ..., \tilde{X}_{(m-1)k+p+2}$

... .

k) $\tilde{X}_k, ..., \tilde{X}_{k+p},\ \tilde{X}_{2k}, ..., \tilde{X}_{2k+p},\ ..., \tilde{X}_{mk}, ..., \tilde{X}_{mk+p}$

Způsob konstrukce odhadu na základě pozorování $(\tilde{X}_1, ..., \tilde{X}_n)$ je následující:

1) Můžeme předpokládat, že $(\tilde{X}_1, ..., \tilde{X}_{p+1})$, $(\tilde{X}_{k+1}, ..., \tilde{X}_{k+p+1})$,..., $(\tilde{X}_{mk+1}, ..., \tilde{X}_{mk+p+1})$ jsou výběry stejně rozdělených nezávislých p-dimensionálních náhodných vektorů. Nalezneme takové p_{ij}, abychom maximalizovali aproximaci věrohodnosti $\sum_{ij} n_{ij}^{(1)} \ln p_{ij}$. Získaný AMLE parametrů $(c, \boldsymbol{\phi}, \sigma^2)$ označíme $(\hat{c}_1, \hat{\boldsymbol{\phi}}_1, \hat{\sigma}_1^2)$.

2) Podobně zkonstruujeme odhad v j-té podskupině a AMLE $(c, \boldsymbol{\phi}, \sigma^2)$ označíme $(\hat{c}_j, \hat{\boldsymbol{\phi}}_j, \hat{\sigma}_j^2), j = 2, ..., k$.

3) Konečné odhady parametrů získáme zprůměrováním dílčích odhadů:

$$\hat{c} = \frac{1}{k}\sum_{i=1}^{k} \hat{c}_i,$$
$$\hat{\boldsymbol{\phi}} = \frac{1}{k}\sum_{i=1}^{k} \hat{\boldsymbol{\phi}}_i,$$
$$\hat{\sigma}^2 = \frac{1}{k}\sum_{i=1}^{k} \hat{\sigma}_i^2.$$

Poznámka: Konzistenci a asymptotickou normalitu uvedených odhadů, které uvádí Bai a kol. (2009), lze dokázat podobným způsobem jako věty 3.3 a 3.4. Pro $p \geq 2$ je výpočet AMLE časově velmi náročný.

Guo a Li (2012) uvádějí obecný AR(p) model ve tvaru

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + \phi_2(X_{t-2} - \mu) + ... + \phi_p(X_{t-p} - \mu) + \varepsilon_t,$$

kde ε_t jsou nezávislé stejně rozdělené veličiny z $N(0, \sigma^2)$ pro $t = 1, ..., n$ a parametr $\boldsymbol{\theta} = (\mu, \boldsymbol{\phi}, \sigma^2), \boldsymbol{\phi} = (\phi_1, ..., \phi_p)$. Oproti zápisu AR($p$) modelu, který používá Bai a kol. (2009), jsou zde veličiny X_i posunuty o jejich střední hodnotu, čímž střední hodnota veličin $X_i - \mu$ bude nula.

Skutečné hodnoty náhodného výběru jsou $\boldsymbol{X} = (X_1, ..., X_n)$, ale k dispozici jsou opět pouze zaokrouhlené hodnoty $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_n)$.

Přitom platí, že $\tilde{X}_t = \tilde{x}_t$ právě když $X_t = x_t$ a $\tilde{x}_t - h/2 \leq x_t < \tilde{x}_t + h/2, t = 1, ..., n$, kde h je šířka intervalu, na který se zaokrouhluje. Platí $\tilde{X}_t = X_t + U_t$, kde U_t má rovnoměrné rozdělení na $[-h/2; h/2]$.

Nechť náhodné veličiny X_i pocházejí ze spojitého rozdělení s hustotou $f(x, \theta)$.

Guo a Li (2012) uvádějí, že maximální věrohodnost pro zaokrohlená data je daná vzorcem $L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) = h^{-n} \int_{\tilde{x}_n - h/2}^{\tilde{x}_n + h/2} ... \int_{\tilde{x}_1 - h/2}^{\tilde{x}_1 + h/2} f(\boldsymbol{u}, \boldsymbol{\theta})\, \mathrm{d}u_1 ... \mathrm{d}u_n$.

Maximálně věrohodný odhad $\hat{\boldsymbol{\theta}}$ parametru $\boldsymbol{\theta}$ získáme maximalizováním věrohodnostní funkce $L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})$.

Lindley (1950) pomocí Maclaurinova vzorce odvodil věrohodnostní funkce v $h = 0$ pro distribuční funkci s jednou proměnou x. Tallis (1967) rozšířil Maclaurinův vzorec na vícerozměrný případ

$$L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) = f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) + \frac{h^2}{24}\sum_{t=1}^{n} \frac{\partial^2 f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})}{\partial \tilde{\boldsymbol{x}}_t^2} + O(h^3).$$

Guo a Li (2012) uvádějí, že logaritmickou věrohodnost lze aproximovat v blízkosti $h = 0$ pomocí

$$L(\boldsymbol{\theta}, \tilde{\boldsymbol{x}}) \sim \ln f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}) + \frac{h^2}{24} \sum_{t=1}^{n} \left[\frac{\frac{\partial^2 f(\boldsymbol{y}, \beta)}{\partial \tilde{\boldsymbol{x}}_t^2}}{f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})} \right] + O(h^3),$$

a dále uvádějí, že korigovaný odhad získáme Newton-Raphsonovou metodou z PMLE

$$\hat{\boldsymbol{\theta}}_A = \boldsymbol{\theta}_0^* - \boldsymbol{A}^{-1}\boldsymbol{b},$$

kde

$$\begin{aligned} &A = [a_{ij}], \boldsymbol{b} = [b_j], \\ &a_{ij} = \frac{\partial^2 \ln(L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \approx \left. \frac{\partial^2 \ln(f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0^*}, \\ &b_j = \frac{\partial \ln(L(\tilde{\boldsymbol{x}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_j} \approx \left. \frac{h^2}{24} \frac{\partial}{\partial \boldsymbol{\theta}_j} \sum_{t=1}^{n} \left[\frac{\frac{\partial^2 f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})}{\partial \tilde{\boldsymbol{x}}_t^2}}{f(\tilde{\boldsymbol{x}}, \boldsymbol{\theta})} \right] \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0^*}, \end{aligned}$$

kde $\boldsymbol{\theta}_i$ je i-tá složka $\boldsymbol{\theta}$, $i, j = 1, ..., p + 2$.

Pro AR(1) uvádějí Guo a Li (2012) i tvary odhadu korigovaného MLE modelu:

$$\begin{aligned} \hat{\mu}_A &= \hat{\mu}_0, \\ \hat{\phi}_A &= \hat{\phi}_0 + h^2 \hat{\phi}_0 \frac{(1 - \hat{\phi}_0^2)}{12\hat{\sigma}_0^2}, \\ \hat{\sigma}_A^2 &= \hat{\sigma}_0^2 - h^2 \frac{(1 + \hat{\phi}_0^2)}{12}, \end{aligned}$$

kde $(\hat{\mu}_0, \hat{\phi}_0, \hat{\sigma}_0^2)$ je PMLE (μ, ϕ_1, σ^2).

Stručný souhrn Gua a Li (2012) vychází z práce Stama a Coggera (1993), kteří se zabývali zaokrouhlenými daty v gaussovských autoregresních řadách.

Stam a Cogger (1993) pracují s AR(p) modelem v centrovaném tvaru

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + ... + \phi_p(X_{t-p} - \mu) + \varepsilon_t, \qquad t = p + 1, ..., n,$$

kde chyby ε_t jsou nezávislé, stejně rozdělené z N(0,σ^2). Odhadovaný parametr je $\boldsymbol{\theta} = (\mu, \phi_1, ..., \phi_p, \sigma^2)^T$. Stam a Cogger (1993) uvádějí věrohodnostní funkci pro

přesná data. Její tvar je

$$L_0(\boldsymbol{X},\boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-n/2} |\boldsymbol{V}_p|^{1/2} \exp\Bigg\{ -(\frac{1}{2\sigma^2})\Bigg[\sum_{i=1}^{p}\sum_{j=1}^{p} v_{ij}(X_i-\mu)(X_j-\mu) + \sum_{t=p+1}^{n}(X_t-\mu-\phi_1(X_{t-1}-\mu)-...-\phi_p(X_{t-p}-\mu))^2\Bigg]\Bigg\}$$

a zavádějí pro ni označení $f(\boldsymbol{X},\boldsymbol{\theta})$.

V uvedeném vzorci je $\sigma^2\boldsymbol{V}_p^{-1}$ kovarianční matice a v_{ij} jsou její prvky. Maximálně věrohodný odhad j-tého prvku parametru $\boldsymbol{\theta}$ získáme řešením systému rovnic

$$\left.\frac{\partial \ln(L_0(\boldsymbol{X},\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0, \qquad j = 1,...,p+2,$$

kde

$$\hat{\boldsymbol{\theta}}_0 = (\hat{\theta}_{1,0}(\boldsymbol{X}),...,\hat{\theta}_{p+2,0}(\boldsymbol{X}))^T.$$

Pro zaokrouhlená data je věrohodnostní funkce dána integrály

$$L_1(\tilde{\boldsymbol{X}},\boldsymbol{\theta}) = \int_{\tilde{x}_n-h/2}^{\tilde{x}_n+h/2} ... \int_{\tilde{x}_1-h/2}^{\tilde{x}_1+h/2} f(\boldsymbol{X},\boldsymbol{\theta})\,\mathrm{d}X_1...\,\mathrm{d}X_n,$$

kde $\tilde{\boldsymbol{X}}$ je sloupcový vektor zaokrouhlených veličin.

Pro j-tý prvek parametru získáme odhad

$$\left.\frac{\partial \ln(L_1(\tilde{\boldsymbol{X}},\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_1(\tilde{\boldsymbol{X}})} = 0, \qquad j = 1,...,p+2.$$

Stam a Cogger (1993) uvádějí, že maximálně věrohodný odhad $\hat{\boldsymbol{\theta}}_0(\boldsymbol{X})$ nelze vypočítat, protože nejsou známa přesná data. Po vzoru Lindleyho (1950) odvozují $\hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}})$. Odhady $\hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}})$ jsou brány jako výchozí bod pro použití Newtonovy metody k získání korigovaného odhadu

$$\hat{\boldsymbol{\theta}}_A(\tilde{\boldsymbol{X}}) = \hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}}) - \boldsymbol{A}^{-1}\boldsymbol{b}$$

kde $\boldsymbol{A}$ je matice s prvky

$$a_{ij} = \frac{\partial^2 \ln(L_0(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j},$$

a $\boldsymbol{b}$ vektor s prvky

$$b_j = \frac{\partial \ln(L_0(\boldsymbol{\theta}, \tilde{\boldsymbol{X}}))}{\partial \boldsymbol{\theta}_j}$$

se známými hodnotami v $\hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}})$.

Analytické vyjádření $\boldsymbol{A}$ a $\boldsymbol{b}$ lze odvodit Taylorovou řadou. Odvození uvádí Lindley (1950) a má tvar

$$L_0(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) = h'' \left[f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) + \frac{h^2}{24} \sum_{t=1}^{n} \frac{\partial f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})}{\partial \tilde{X}_t^2} \right] + O(h^3).$$

Úpravou a zlogaritmováním dostaneme

$$\ln(L_0(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})) = n \ln(h) + \ln(f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})) + \ln \left[1 + \frac{h^2}{24} \sum_{t=1}^{n} \left[\frac{\partial f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})}{\partial \tilde{X}_t^2} / f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) \right] + O(h^3) \right]$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln(L_0(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})) = \frac{\partial}{\partial \boldsymbol{\theta}_i} \ln(f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})) + \frac{h^2}{24} \frac{\partial}{\partial \boldsymbol{\theta}_i} \sum_{t=1}^{n} \left[\frac{\partial f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})}{\partial \tilde{X}_t^2} / f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) \right] + O(h^3),$$
$$i = 1, ..., p+2.$$

Při použití dostáváme v bodě $\hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}})$ přibližné vyjádření pro elementy

$$\begin{aligned} a_{ij} &\doteq \frac{\partial^2 \ln(f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j} \bigg|_{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}}) \\ &= \hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}), \qquad i, j = 1, ..., p+2 \end{aligned}$$

$$\begin{aligned} b_j &\doteq \frac{h^2}{24} \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \left[\frac{\partial f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta})}{\partial \tilde{X}_i^2} / f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}) \right] \bigg|_{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0(\tilde{\boldsymbol{X}}) \\ &= \hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}), \qquad j = 1, ..., p+2 \end{aligned}$$

Přestože uvedené vzorce vypadají složitě, jsou snadno odvoditelné a ve specifických aplikacích je lze vyjádřit numericky. Pro výběry o velkých velikostech (vysoké n) je vhodné zanedbat ty členy, které jsou vzhledem k n velmi malé.

Kapitola 4

Ukázky vlivu zaokrouhlených dat na různých modelech

4.1 Lineární regrese

Mějme náhodný výběr $X_1, ..., X_n$ z normálního rozdělení se střední hodnotou μ a rozptylem σ^2. Uvažujme lineární model $\boldsymbol{Y} = \beta_0\mathbf{1} + \beta_1\boldsymbol{X} + \boldsymbol{\varepsilon}$, kde $\boldsymbol{X} = (X_1, ..., X_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^T$. Chyby $\varepsilon_1, ..., \varepsilon_n$ pocházejí z normálního rozdělení se střední hodnotou μ_ε a rozptylem σ_ε^2.

Pro vyšetřování chování na nasimulovaných datech budeme nejprve uvažovat případ, kdy $n = 100$, $\mu = 0$, $\sigma_x^2 = 25$, $\mu_\varepsilon = 0$, $\sigma_\varepsilon^2 = 25$, $\beta_0 = 4, \beta_1 = 3$.

Náhodný výběr $X_1, ..., X_n$ budeme postupně zaokrouhlovat. Nejdříve hodnoty z náhodného výběru zaokrouhlíme na 4 desetinná místa. Dále budeme napozorované hodnoty zaokrouhlovat hruběji a budeme zkoumat vliv zaokrouhlení na odhad parametrů β_0 a β_1 pořízený metodou nejmenších čtverců. Odhady jsou shrnuté v tabulce 4.1 .

Tabulka 4.1: Porovnání odhadů parametrů MNČ na základě zaokrouhlených dat

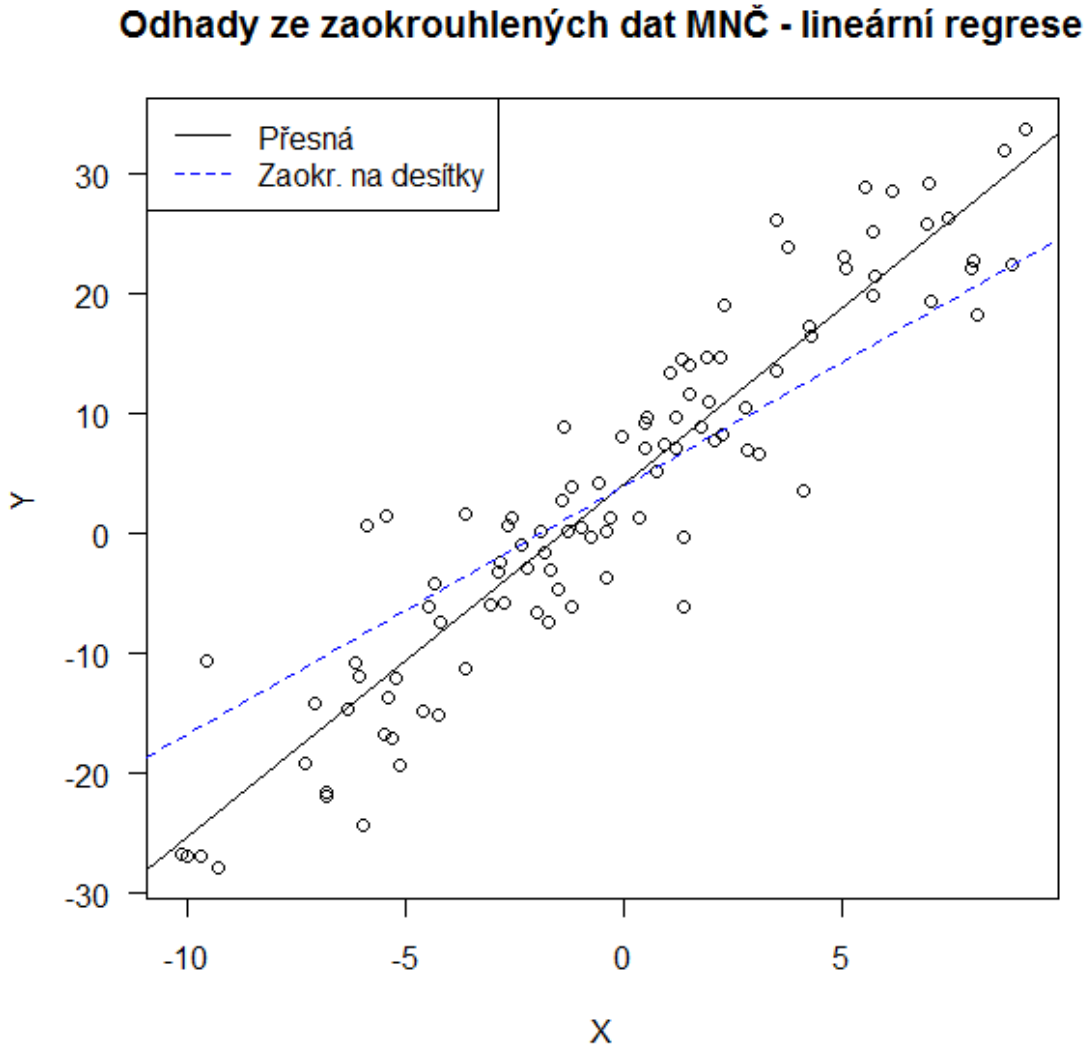
Zaokrouhlení	Bodový odhad β_0	Bodový odhad β_1
žádné	4,1402	2,9446
desetitisíciny	4,1402	2,9446
tisíciny	4,1401	2,9446
setiny	4,1395	2,9446
desetiny	4,1314	2,9462
jednotky	4,0837	2,9443
desítky	3,8513	2,0714

Odhady získané metodou nejmenších čtverců na základě dat zaokrouhlených dle uvedené hrubosti. Odhady jsou učiněny na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 4$, $\beta_1 = 3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 25.

Po vykreslení přímek je na grafu 4.1 vidět, že od přímky s parametry odhadnutými z nezaokrouhlených pozorování je rozlišitelná pouze přímka s parametry

odhadnutými z pozorování zaokrouhlenými na desítky.

Obrázek 4.1: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



Odchylka přímky s parametry odhadnutými MNČ na základě dat zaokrouhlených na desítky od skutečné přímky, nasimulované pomocí $\boldsymbol{X}$ a $\boldsymbol{Y}$. Odhad je učiněn na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 4$, $\beta_1 = 3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 25.

Pokud pracujeme se zaokrouhlenými daty, je vhodnější namísto metody nejmenších čtverců odhadnout parametry přímky pomocí momentové metody pro strukturální relaci (model s nepřesnými hodnotami, ve kterém nemůže experimentátor zvolit hodnoty x_i).

Uvažujme dvojice veličin (x_i, y_i), mezi nimiž platí lineární vztah $y_i = \beta_0 + \beta_1 x_i, i = 1, ..., n$. Namísto dvojic (x_i, y_i) můžeme pozorovat pouze dvojice $(\tilde{x}_i, \tilde{y}_i)$, které jsou zatíženy náhodnými chybami.

Budeme předpokládat, že $\tilde{x}_i = x_i + u_i$, $\tilde{y}_i = y_i + \gamma_i$, $\quad i = 1, ..., n$, kde

$u_i, i = 1, ..., n$ pochází z rovnoměrného rozdělení $R(a, b)$ a $\gamma_i, i = 1, ..., n$ pochází z normálního rozdělení $N(0, \sigma_\gamma^2)$.

Dále budeme předpokládat, že chyby u_i a γ_i jsou vzájemně nekorelované a skutečné hodnoty x_i jsou nekorelované s oběma těmito chybami (u_i i γ_i).

Předpokládáme, že $x_1, ..., x_n$ jsou nezávislé náhodné veličiny s normálním rozdělením $N(\mu, \sigma_x^2)$. Označení malými písmeny má vyjadřovat obecný lineární model, kde jsou veličiny zatíženy chybami, nikoliv pouze jeden konkrétní vztahující se k náhodnému výběru, který jsme získali simulací.

Po dosazení do lineárního vztahu namísto y_i a x_i dostáváme:

$$\tilde{y}_i = \beta_0\beta_1\tilde{x}_i + (\gamma_i - \beta_1 u_i), i = 1, ..., n.$$

Jelikož

$$\operatorname{cov}(\tilde{x}_i, \gamma_i - \beta_1 u_i) = \operatorname{cov}(x_i + u_i, \gamma_i - \beta_1 u_i) = -\beta_1\sigma_u^2$$

není v obecném případě nula, hodnoty nezávisle proměnné závisejí na vektoru chyb. Pak

$$\begin{aligned}
E\tilde{x}_i &= E(x_i + \delta_i) = Ex_i + E\delta_i = \mu + 0 = \mu, \\
E\tilde{y}_i &= E(y_i + \gamma_i) = E\tilde{y} = y_i + E\gamma_i = \beta_0 + \beta_1\mu + 0 = \beta_0 + \beta_1\mu \\
\operatorname{cov}(\tilde{x}_i, \tilde{y}_i) &= \operatorname{cov}(x_i + \delta_i, y_i + \gamma_i) = \operatorname{cov}(x_i + \delta_i, \beta_0 + \beta_1 x_i + \gamma_i) \\
&= \operatorname{cov}(x_i + \delta_i, \beta_1 x_i + \gamma_i) = \beta_1\sigma_x^2, \\
\operatorname{var}(\tilde{x}_i) &= \operatorname{var}(x_i + \delta_i) = \operatorname{cov}(x_i + \delta_i, x_i + \delta_i) = \sigma_x^2 + \sigma_\delta^2 \\
\operatorname{var}(\tilde{y}_i) &= \operatorname{var}(y_i + \gamma_i) = \operatorname{var}(\beta_0 + \beta_1 x_i + \varepsilon_i) \\
&= \operatorname{cov}(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_1^2\sigma_x^2 + \sigma_\varepsilon^2.
\end{aligned}$$

Odhady parametrů získáme momentovou metodou.

Dostáváme:

$$\begin{aligned}
\mu &= \bar{\tilde{x}}, \\
\beta_0 + \beta_1\mu &= \bar{\tilde{y}}, \\
\beta_1\sigma_x^2 &= s_{\tilde{x},\tilde{y}}, \\
\sigma_x^2 + \sigma_\delta^2 &= s_{\tilde{x}}^2, \\
\beta_1^2\sigma_x^2 + \sigma_\gamma^2 &= s_{\tilde{y}}^2,
\end{aligned}$$

kde $\bar{\tilde{x}} = \frac{1}{n}\sum_{i=1}^n \tilde{x}_i$ je průměr, $s_{\tilde{x},\tilde{y}} = \frac{1}{n}\sum_{i=1}^n \left((\tilde{x}_i - \bar{\tilde{x}})(\tilde{y}_i - \bar{\tilde{y}})\right)$ je výběrová kovariance a $s_{\tilde{x}}^2 = s_{\tilde{x},\tilde{x}}$ je výběrový rozptyl. To je pět rovnic pro šest neznámých.

Ve výše uvedeném případě můžeme předpokládat, že při zaokrouhlení $X_i, i = 1, ..., 100$, na desítky má δ rovnoměrné rozdělení na intervalu $(-5, 5)$.

Tento předpoklad otestujeme pomocí Kolmogorovova-Smirnovova testu proti oboustranné alternativě.

Tabulka 4.2: Testování hypotézy, že chyby vzniklé zaokrouhlením dat pocházejí z rovnoměrného rozdělení

Zaokrouhlení	p-hodnota
desetitisíciny	0,6253
tisíciny	0,6583
setiny	0,07386
desetiny	0,4421
jednotky	0,8989
desítky	0,93

Výsledky testování hypotéz, že chyby u_i vzniklé zaokrouhlením dat X_i pocházejí z rovnoměrných rozdělení s parametry závislými na hrubosti zaokrouhlení (z R($-0,00005$;-0,00005) pro zaokrouhlení na desetitisíciny ,..., z R(-5,5) pro zaokrouhlení na desítky). Hypotézy byly testovány proti oboustranné alternativě pomocí Kolmogorovova-Smirnovova testu.

Výsledky testu (shrnuté v tabulce 4.2) nejsou na hladině $\alpha = 0,05$ v rozporu s předpokladem, že chyby zaokrouhlování pocházejí z rovnoměrného rozdělení.

Z očekávaného rozdělení u dokážeme určit jeho rozptyl — rozptyl rovnoměrného rozdělení na intervalu (a, b) je $\frac{(b-a)^2}{12}$, čímž v našem případě při zaokrouhlení na desítky dostaneme $\sigma_u^2 = \frac{100}{12} = 8,333...$.

Tím dostáváme pět rovnic pro pět neznámých. Jejich vyřešením z náhodného výběru $x_1, ..., x_{100}$ zaokrouhleného na desítky (tedy $\tilde{x}_1, ..., \tilde{x}_{100}$) dostaneme odhady koeficientů lineárního modelu $\hat{\beta}_0 = 4,032524$, $\hat{\beta}_1 = 2,675403$.

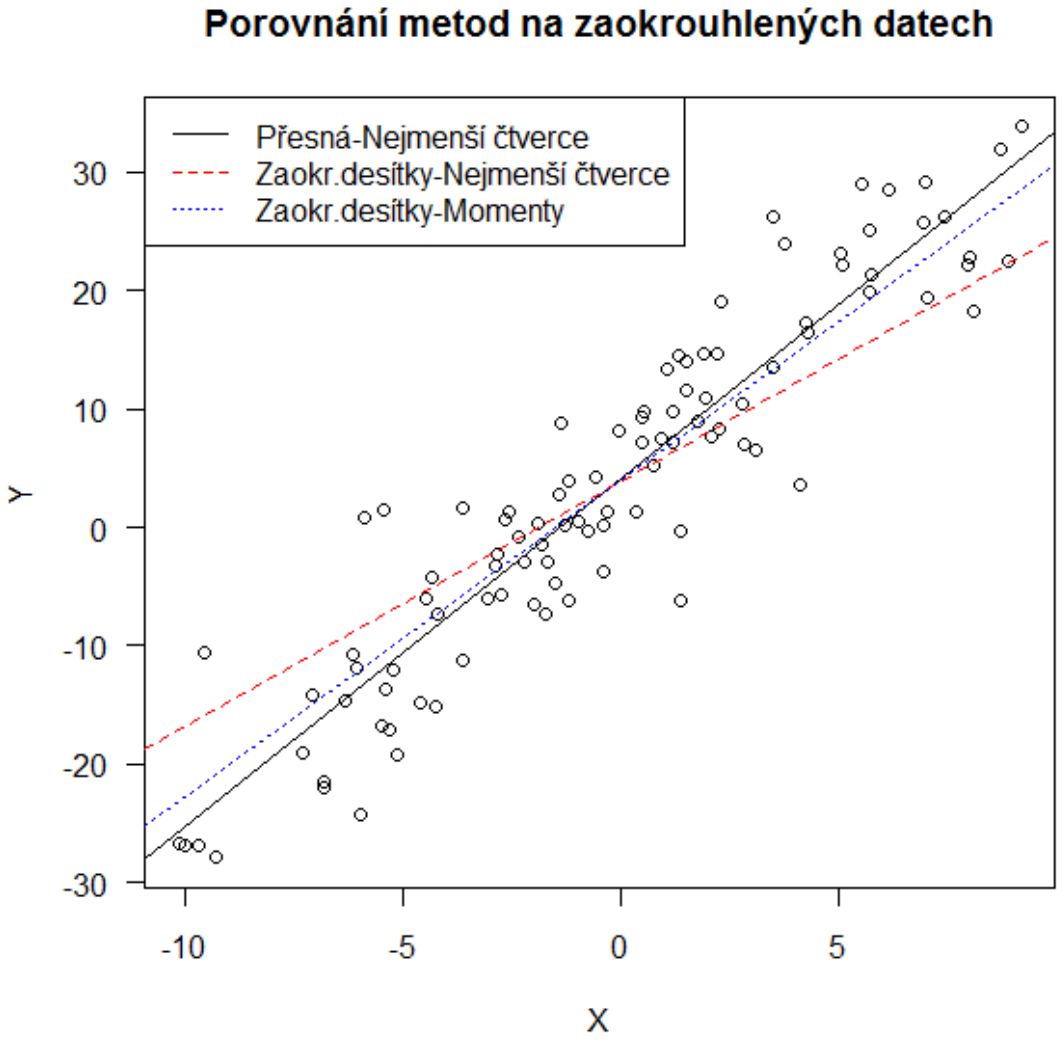
Na grafu 4.2 je názorně vidět, že na datech, kde se v odhadech parametrů zaokrouhlení projeví, funguje odhad momentovou metodou lépe než odhad metodou nejmenších čtverců.

Dále vyšetříme, zda se vliv zaokrouhlení projeví při méně hrubém zaokrouhlení, pokud řádově zmenšíme parametry. Budeme nyní uvažovat lineární model s parametry $\beta_0 = 0,4$, $\beta_1 = 0,3$. Pomocí metody nejmenších čtverců dostaneme odhady shrnuté v tabulce 4.3. Tabulka 4.3 se od tabulky 4.1 liší tím, že parametry lineární regrese jsou 0,4 a 0,3, nikoliv 4 a 3.

Po vykreslení grafu 4.3 se od přímky vypočítané z přesných dat liší pouze přímka s parametry odhadnutými na základě dat zaokrouhlených na desítky. Na grafu 4.3 je také vidět rozmístění bodů okolo přímky — body jsou více rozptýleny než v předchozím případě, neshlukují se blízko přímky. Jedním z možných vysvětlení je výraznější vliv chybového vektoru pro $|\beta_1| < 1$ v případě, že náhodný výběr $X_1, ..., X_{100}$ i chyby $\varepsilon_i, i = 1, ..., 100$, pocházejí z normálního rozdělení se stejným rozptylem.

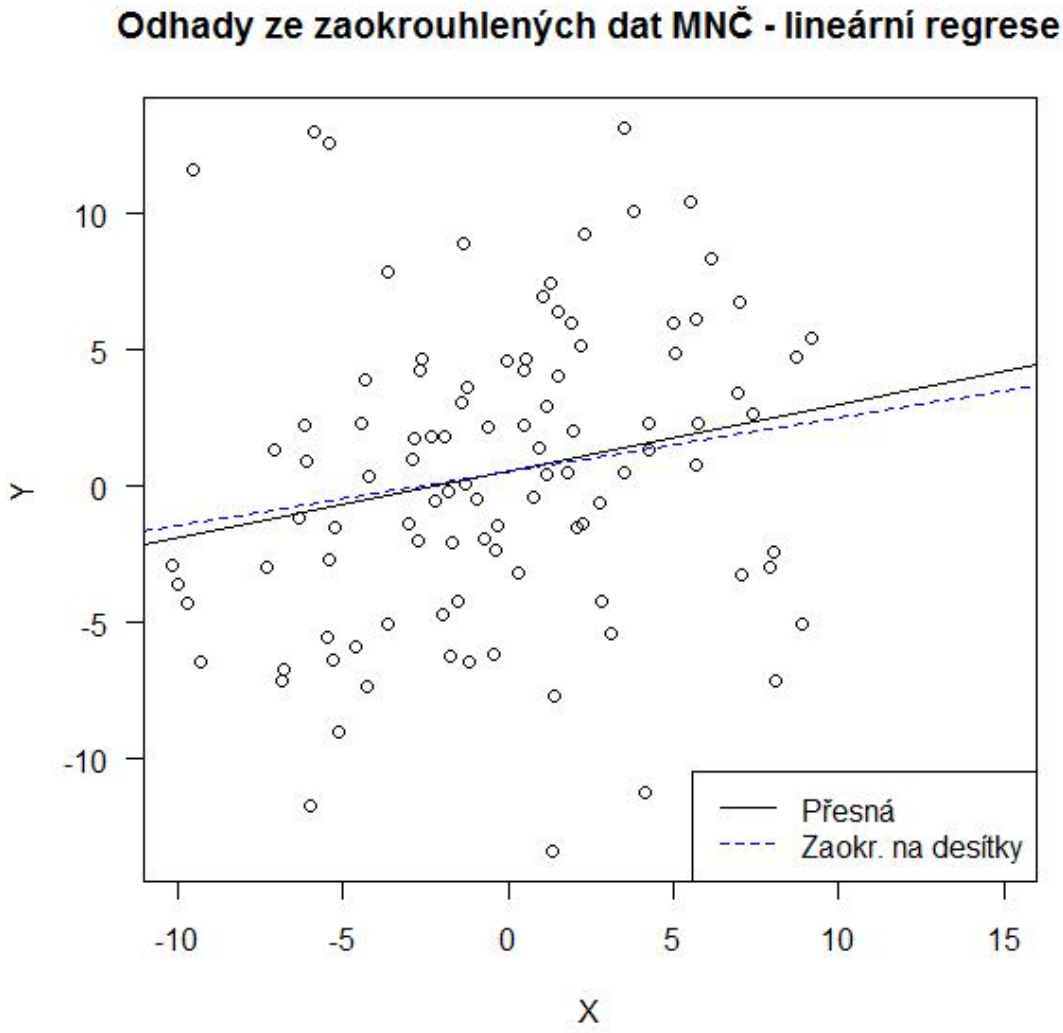
Dále porovnáme vliv zaokrouhlení pro lineární model s parametry $\beta_0 = 4$, $\beta_1 = 3$, kde náhodná veličina X i chyby ε mají normální rozdělení se střední

Obrázek 4.2: Porovnání metody nejmenších čtverců a momentové metody pro zaokrouhlená data



Porovnání odchylky přímky s parametry odhadnutými metodou nejmenších čtverců a přímky s parametry odhadnutými momentovou metodou (z dat zaokrouhlených na desítky) od skutečné přímky.

Obrázek 4.3: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



Odchylka přímky s parametry odhadnutými MNČ na základě dat zaokrouhlených na jednotky od skutečné přímky, nasimulované pomocí $\boldsymbol{X}$ a $\boldsymbol{Y}$. Odhad je učiněn na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 0,4 \quad$, $\beta_1 = 0,3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 25.

Tabulka 4.3: Porovnání odhadů parametrů MNČ na základě zaokrouhlených dat

Zaokrouhlení	Bodový odhad β_0	Bodový odhad β_1
žádné	0,54020	0,24460
desetitisíciny	0,54020	0,24460
tisíciny	0,54020	0,24460
setiny	0,54010	0,24460
desetiny	0,53940	0,24470
jednotky	0,53520	0,24370
desítky	0,52382	0,19757

Odhady získané metodou nejmenších čtverců na základě dat zaokrouhlených dle uvedené hrubosti. Odhady jsou učiněny na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 0,4 \quad , \beta_1 = 0,3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 25.

Tabulka 4.4: Porovnání odhadů parametrů MNČ na základě zaokrouhlených dat

Zaokrouhlení	Bodový odhad β_0	Bodový odhad β_1
žádné	4,20101	3,03140
desetitisíciny	4,20101	3,03140
tisíciny	4,20097	3,03139
setiny	4,20071	3,03148
desetiny	4,17823	3,02947
jednotky	4,17650	2,79690

Odhady získané metodou nejmenších čtverců na základě dat zaokrouhlených dle uvedené hrubosti. Odhady jsou učiněny na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 4$, $\beta_1 = 3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 1.

hodnotou 0 a rozptylem 1.

Pomocí metody nejmenších čtverců dostaneme odhady shrnuté v tabulce 4.4.

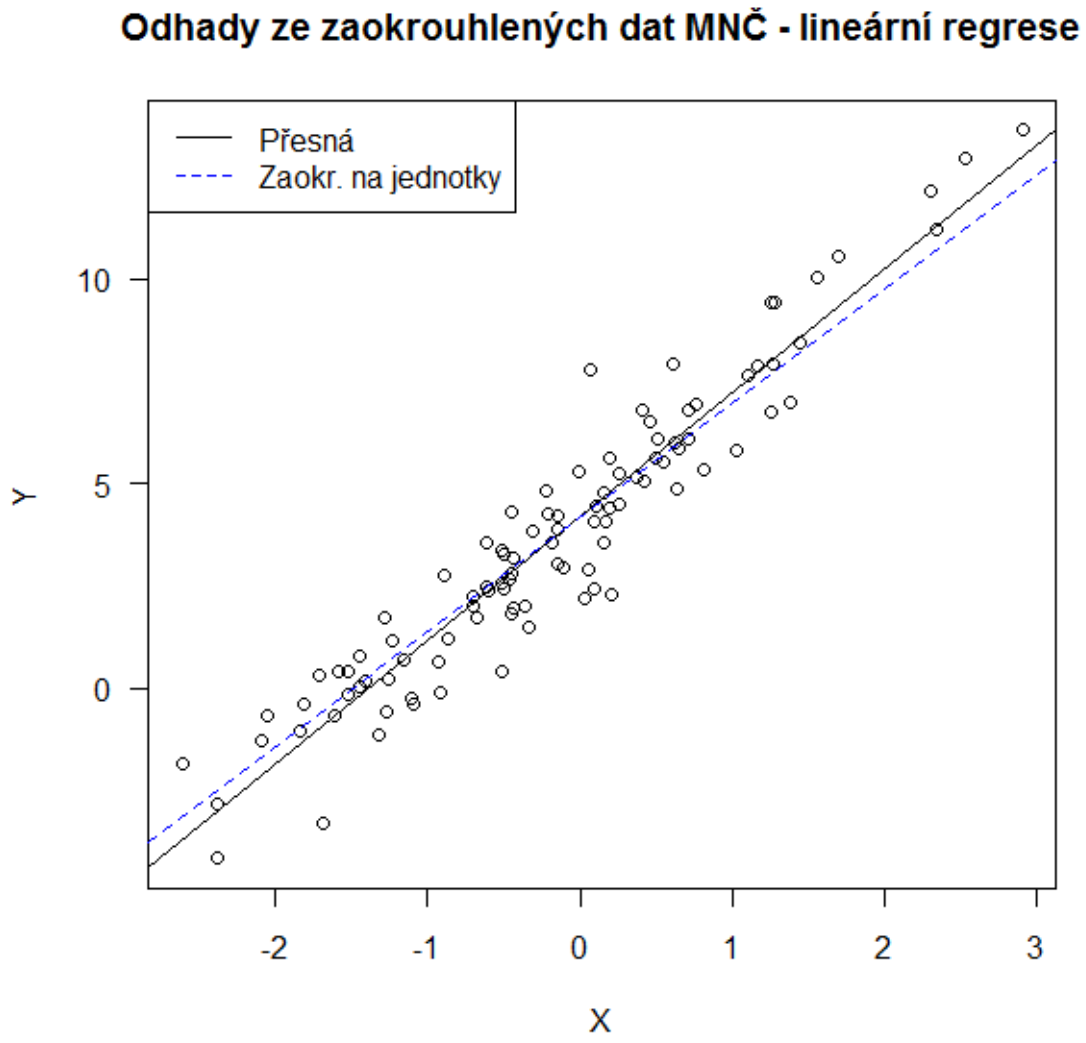
Po vykreslení grafu 4.4 se od přímky s parametry odhadnutými z nezaokrouhlených pozorování viditelně liší pouze přímka s parametry odhadnutými z pozorování zaokrouhlenými na jednotky. Zaokrouhlení na desítky je natolik hrubé, že na základě takových dat nelze odhadovat parametry — všechna data se zaokrouhlí na stejné číslo.

Pomocí Kolmogorovova-Smirnovova testu opět otestujeme hypotézu, že zaokrouhlovací chyby $\boldsymbol{U} = \boldsymbol{X} - \tilde{\boldsymbol{X}}$ pocházejí z rovnoměrného rozdělení, jehož parametry závisí na řádu, na který zaokrouhlujeme. Hypotézu testujeme proti oboustranné alternativě.

Tentokrát na hladině $\alpha = 0,05$ zamítneme nulovou hypotézu u dat zaokrouhlených na desetiny.

Jelikož u zaokrouhlení na jednotky jsme nulovou hypotézu nezamítli, bu-

Obrázek 4.4: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



Odchylka přímky s parametry odhadnutými MNČ na základě dat zaokrouhlených na jednotky od skutečné přímky, nasimulované pomocí $\boldsymbol{X}$ a $\boldsymbol{Y}$. Odhad je učiněn na základě 100 pozorování X_i. Je použit lineární model s parametry $\beta_0 = 4$, $\beta_1 = 3$. Nezávislé veličiny X_i i chyby ε_i mají nulovou střední hodnotu a rozptyl 1.

Tabulka 4.5: Testování hypotézy, že chyby vzniklé zaokrouhlením dat pocházejí z rovnoměrného rozdělení

Zaokrouhlení	p-hodnota
desetitisíciny	0,8326
tisíciny	0,656
setiny	0,8876
desetiny	0,0234
jednotky	0,6726

Výsledky testování hypotéz, že chyby u_i vzniklé zaokrouhlením dat X_i pocházejí z rovnoměrných rozdělení s parametry závislými na hrubosti zaokrouhlení (z R($-0,00005$; 0,00005) pro zaokrouhlení na desetitisíciny ,..., z R(-5;5) pro zaokrouhlení na desítky). Hypotézy byly testovány proti oboustranné alternativě pomocí Kolmogorovova-Smirnovova testu.

deme dále pro použití metody momentů předpokládat, že chyba u pochází z $R(-0,5; 0,5)$, z čehož dostaneme její rozptyl $\sigma_u^2 = \frac{1}{12} = 0,08333....$

Dostáváme odhady parametrů $\hat{\beta}_0 = 4,211866$, $\hat{\beta}_1 = 2,983287$.

Na obrázku 4.5 je opět vidět, že momentová metoda dává lepší výsledek než metoda nejmenších čtverců.

Uvažujme opět dvojice veličin (x_i, y_i), mezi nimiž platí lineární vztah $y_i = \beta_0 + \beta_1 x_i, i = 1, ..., n$. K dispozici máme opět pouze dvojice $(\tilde{x}_i, \tilde{y}_i)$ zatížené náhodnými chybami. Dále předpokládáme, že platí $\tilde{x}_i = x_i + u_i$, $\tilde{y} = y_i + \varepsilon_i, i = 1, ..., n$.

Hodnoty veličin $x_i, i = 1, ..., n$, nebudou tentokrát pocházet z náhodného výběru, ale budou zadány experimentátorem.

Hodnoty $x_i, i = 1, ..., n$, zvolíme ekvidistantně rozdělené na intervalu [20,120]. Jejich počet bude 100, resp. 200. Po přidání náhodných chyb a zaokrouhlení porovnáme odhady pomocí metody nejmenších čtverců a momentové metody.

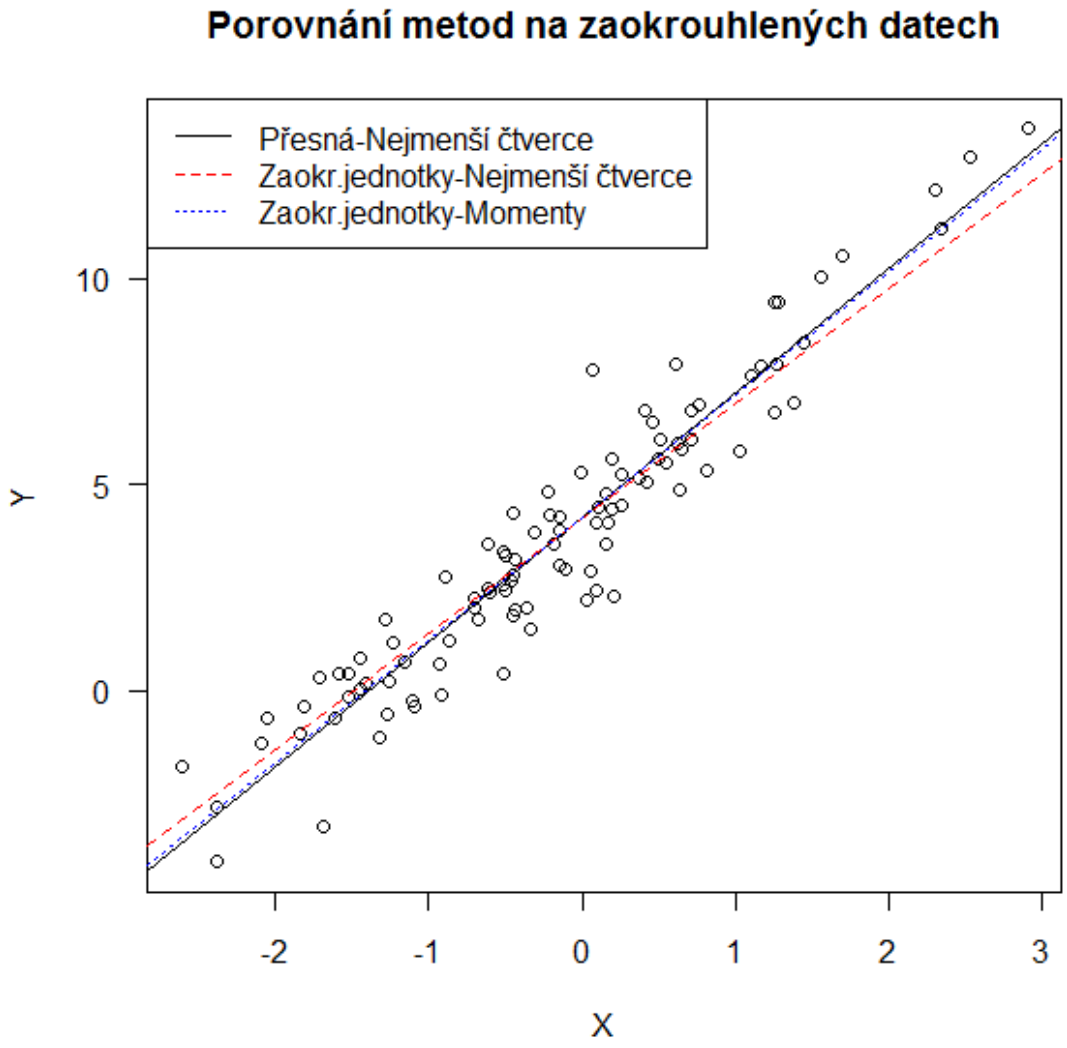
V tabulkách 4.6 a 4.7 jsou shrnuta výsledná data o odhadnutých parametrech pro 1000 pozorování.

V tabulce 4.6 vidíme, že vyjma případů, kdy máme 200 hodnot x_i a rozptyl chyby ε je 25, je střední hodnota odhadu $\hat{\beta}_0$ získaná metodou momentů přesnější, než střední hodnota odhadu $\hat{\beta}_0$ získaná metodou nejmenších čtverců. V případě, že máme 100 hodnot x_i, má metoda momentů ve většině případů i menší rozptyl. Pro 100 hodnot x_i a zaokrouhlování na jednotky dává metoda momentů výrazně lepší výsledek — u metody nejmenších čtverců neleží skutečná hodnota parametru β_0 v intervalu mezi nejmenší a největší hodnotou odhadu.

V tabulce 4.7 vidíme, že pro 100 hodnot x_i dostáváme lepší střední odhad $\hat{\beta}_0$ metodou momentů. Pro 200 hodnot x_i a rozptyl chyby ε o hodnotě 25 dává naopak přesnější střední hodnotu odhadů metoda nejmenších čtverců. Pro 100 hodnot x_i vykazuje metoda momentů mírně menší rozptyl, zatímco pro 200 hodnot x_i ho má převážně větší. U obou metod v některých případech neleží skutečná hodnota parametru mezi nejmenší a největší hodnotou odhadu.

V případě, kdy veličiny x_i nejsou náhodně vybrané, ale zvolené, nemůžeme jednoznačně určit, že metoda momentů dává jednoznačně lepší výsledky, třebaže

Obrázek 4.5: Porovnání metody nejmenších čtverců a momentové metody pro zaokrouhlená data



Porovnání odchylky přímky s parametry odhadnutými metodou nejmenších čtverců a přímky s parametry odhadnutými momentovou metodou (z dat zaokrouhlených na jednotky) od skutečné přímky.

Tabulka 4.6: Porovnání dat odhadů MNČ a MM ze zaokrouhlených dat pro 1000 opakování — parametr β_0

Počet x_i	Metoda	Zaokrouhlení	σ_ε^2	E $\hat{\beta}_0$	$\hat{sd}(\hat{\beta}_0)$	Min $\hat{\beta}_0$	Max $\hat{\beta}_0$
100	MNČ	na jednotky	1	2,985874	0,2482159	2,215446	3,730548
100	MM	na jednotky	1	3,984310	0,2472867	3,217732	4,724988
100	MNČ	na jednotky	25	5,038096	0,2611939	4,131239	5,890064
100	MM	na jednotky	25	3,482064	0,2607174	2,567284	4,329015
100	MNČ	na desítky	1	2,878757	2,6330050	-5,358748	12,875080
100	MM	na desítky	1	3,878682	2,6214260	-4,283932	13,809340
100	MNČ	na desítky	25	5,104095	2,5493990	-2,908127	13,790460
100	MM	na desítky	25	3,554961	2,5578040	-4,439155	12,352060
200	MNČ	na jednotky	1	5,036467	0,1846527	4,470452	5,616086
200	MM	na jednotky	1	3,981151	0,1855177	3,412147	4,563613
200	MNČ	na jednotky	25	3,992973	0,1830850	3,270428	4,574417
200	MM	na jednotky	25	3,477374	0,1826254	2,760031	4,064312
200	MNČ	na desítky	1	4,923054	1,8920740	-1,629820	12,205450
200	MM	na desítky	1	3,868034	1,9003020	-2,738511	11,183170
200	MNČ	na desítky	25	3,955123	1,8016360	-2,660857	9,955850
200	MM	na desítky	25	3,430661	1,8010370	-2,836436	9,445779

Porovnání vlastností odhadů parametru β_0 z modelu $\tilde{\boldsymbol{Y}} = \beta_0 + \beta_1 \tilde{\boldsymbol{X}} + \boldsymbol{\varepsilon}$ pomocí metody nejmenších čtverců a momentové metody. Data $\boldsymbol{X}$ ekvidistantně rozdělená z intervalu [20 ;120] jsou zaokrouhlená na uvedenou hrubost, dodatečné chyby $\boldsymbol{\varepsilon}$ pocházejí z normálního rozdělení s nulovou střední hodnotou a uvedeným rozptylem.

na námi zvolených datech vykazuje nedostatky o trochu menší než metoda nejmenších čtverců.

Tabulka 4.7: Porovnání dat odhadů MNČ a MM ze zaokrouhlených dat zaokrouhlených dat pro 1000 opakování - parametr β_1

Počet x_i	Metoda	Zaokrouhlení	σ_ε^2	E $\hat{\beta}_1$	$\hat{sd}(\hat{\beta}_1)$	Min $\hat{\beta}_1$	Max $\hat{\beta}_1$
100	MNČ	na jednotky	1	3,014520	0,003244488	3,005213	3,024058
100	MM	na jednotky	1	3,000256	0,003229787	2,991006	3,009739
100	MNČ	na jednotky	25	2,985196	0,003457454	2,970352	2,995430
100	MM	na jednotky	25	3,007425	0,003461356	2,992653	3,017918
100	MNČ	na desítky	1	3,015219	0,034448460	2,888962	3,118973
100	MM	na desítky	1	3,000934	0,034271010	2,875616	3,103713
100	MNČ	na desítky	25	2,984835	0,033841190	2,883700	3,097253
100	MM	na desítky	25	3,006965	0,034034590	2,908589	3,119562
200	MNČ	na jednotky	1	2,985225	0,002501876	2,977729	2,993112
200	MM	na jednotky	1	3,000301	0,002514906	2,992777	3,008230
200	MNČ	na jednotky	25	3,000074	0,002374086	2,993228	3,009242
200	MM	na jednotky	25	3,007440	0,002368791	3,000364	3,016534
200	MNČ	na desítky	1	2,986158	0,025448210	2,884145	3,079006
200	MM	na desítky	1	3,001230	0,025576400	2,898749	3,094845
200	MNČ	na desítky	25	3,001127	0,023579720	2,932489	3,080787
200	MM	na desítky	25	3,008619	0,023532760	2,934066	3,083295

Porovnání vlastností odhadů parametru β_1 z modelu $\tilde{\boldsymbol{Y}} = \beta_0 + \beta_1 \tilde{\boldsymbol{X}} + \boldsymbol{\varepsilon}$ pomocí metody nejmenších čtverců a momentové metody. Data $\boldsymbol{X}$ ekvidistantně rozdělená z intervalu [20 ;120] jsou zaokrouhlená na uvedenou hrubost, dodatečné chyby $\boldsymbol{\varepsilon}$ pocházejí z normálního rozdělení s nulovou střední hodnotou a uvedeným rozptylem.

Kapitola 5

Závěr

V práci jsme shrnuli metody pro výpočet parametrů ze zaokrouhlených dat na MA a AR modelech časových řad. Pomocí experimentu jsme zkoumali, zda se zaokrouhlená data AR(1) modelu chovají jako data modelu ARMA(1,1) - na použité velikosti výběru a jemnosti zaokrouhlení se toto chování neprokázalo. Dále jsme pomocí experimentu porovnávali metodu nejmenších čtverců a momentovou metodu pro lineární regresi na zaokrouhlených datech. Jednoznačně lepší výsledky vykazovala momentová metoda na náhodných datech při hrubém zaokrouhlení. V ostatních případech se nepotvrdilo, že by byla výrazně lepší než metoda nejmenších čtverců.

Zjistili jsme, že na náhodných výběrech o 1000 veličinách se jemné zaokrouhlení výrazně neprojevuje.

Literatura

[1] Anděl J. (1976): Statistická analýza časových řad. SNTL - Nakladatelství technické literatury, Praha.

[2] Anděl J. (2007a): Statistické metody. Matfyzpress, Praha.

[3] Anděl J. (2007b): Základy matematické statistiky. Matfyzpress, Praha.

[4] Bai Z., Zheng S., Zhang B., Hu G. (2009): Statistical analysis for rounded data. J. Statist. Planning Infer. 139, 2526-2542.

[5] Brockwell P.J., Davis R.A. (1991): Time Series: Theory and Methods. Springer-Verlang, New York.

[6] Dempster A.P., Rubin D.B (1983): Rounding error in regression: The appropriatness of Sheppard's connections. Journal of teh Royal Statistical Society, Ser. B, 45, 51-59.

[7] Guo M., Li G.-L. (2012): Estimation of MA(1) model base on rounded data. Tatra Mountains 51, 45-53.

[8] Jarník V. (1974): Diferenciální počet (I). Academia, nakladatelství Československé akademie věd, Praha.

[9] Lachout P. (2008): Matematické programování. Pracovní text k přednášce EKN011, Optimalizace I.

[10] McLeod A.I. , Zhang Y. (2006). Partial Autocorrelation Parameterization for Subset Autoregression. Journal of Time Series Analysis, 27, 599-612.

[11] McLeod A. I. , Zhang Y. (2007). Faster ARMA maximum likelihood estimation, Computational Statistics & Data Analysis 52(4), URL http://dx.doi.org/10.1016/j.csda.2007.07.020

[12] Lindley D. V. (1950): Grouping corrections and maximum likelihood equations. P. Camb. Philos. Soc. 46, 106-110.

[13] Prášková Z. (2004): Základy náhodných procesů II. Karolinum, Praha.

[14] Prokešová M. (2011): Základy matematického modelování.

[15] Sheppard W.F. (1898): On the calculation of the most probable values of frequency constantsfor data arranged according to equidistant divisons of a scale. Proceedings of the Lonon Mathematical Society, 29, 353-380.

[16] Stam A., Cogger K. O. (1993): Rounding errors in autoregressive processes. Internat. J. Forecast. 9, 487-508.

[17] Tallis G. M. (1967): Approximate maximum likelihood estimates from grouped data. Technometric 9, 599-606.

[18] Tricker A. (1990a): The effect of rounding on the significance level of certain normal test statistics. J. Appl. Statist. 17, 31-38.

[19] Tricker A. (1990b): The effect of rounding on the power level of certain normal test statistics. J. Appl. Statist. 17, 219-228.

[20] www.wolframalpha.com