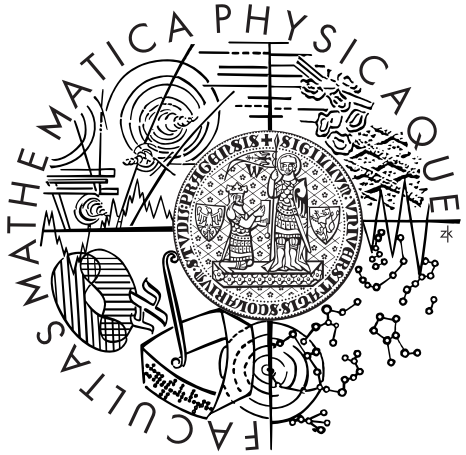
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

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: Pravděpodobnost a Matematická statistika
Studijní obor: Matematická statistika

Praha 2014

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 disusses 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 adition, the comparison of linear regression is shown at graphs.

Keywords: rounded data, time series, linear regression

Obsah

Kapitola 1

Úvod

Tato práce bude pojednávat 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 na zaokrouhlených datech 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 Metoda maximální věrohodnosti

2.1.1 Metoda maximální věrohodnosti pro přesná data

Použití metody maximální věrohodnosti předpokládá, že známe typ rozdělení, ze kterého náhodný výběr pochází. Uvažujme spojité rozdělení s hustotou $f(x, \theta)$, 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), jeho maximálně věrohodný odhad zpravidla získáme řešením soustavy

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

2.1.2 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.

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

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

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 $p(nh,\theta)$ vyjádřuje pravděpodobnost, že zaokrouhlená veličina leží v příslušném intervalu, po převodu ze spojitého případu na diskrétní s ní nahradíme původní spojitou hustotu.

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

Chceme vypočítat hodnotu θ_1^*, ale máme vypočtenu 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.

Dále podrobněji rozvedeme odhad pro Δ, který uvádí Lindley (1950). Předpokládejme, že hustota f má třetí derivaci. Pro snazší úpravy rozvineme funkci f v bodě nh pomocí Taylorova polynomu. Při použití Taylorova rozvoje vyššího řádu bychom získali jemnější aproximaci. 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 derivace funkce f.

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 vyintegrované výrazy a z nich odvodili korekci pro maximálně věrohodný odhad ze zaokrouhlených dat. Postupně dostaneme

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

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

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

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

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

kde R' je zbytkový člen obsahující derivaci třetího řádu funkce $f(nh, \theta)$. To vychází z Taylorova rozvoje, kterým jsme funkci f rozvíjeli do nejvýše druhého řádu a zbytek jsme zahrnuli do zbytkového členu R.

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

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

Derivováním podle θ dostáváme

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

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

Jelikož poslední dva členy obsahují 3. nebo vyšší mocninu h, můžeme je shrnout pod výraz $O(h^3)$. Tím 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 s počáteční hodnotou parametru θ_0^*, vypočítaného z rovnice, jako by šlo o nezaokrouhlená data s prvotní aproximací $\sum_{i=1}^{k}\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f(n_i h,\theta_0) = 0$.

Jako výslednou korekci Lindley (1950) uvádí

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

2.2 Momentová metoda

Momentová metoda podobně jako metoda maximální věrohodnosti 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 $N(\mu, \sigma_x^2)$. 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.

2.3 Metoda nejmenších čtverců

Metoda nejmenších čtverců 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 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 $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z^T Z})^{-1}\boldsymbol{Z^T Y}$. 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.

Kapitola 3

Aplikace na modely časových řad

3.1 Model MA

Bai a kol. (2009) uvádějí odhady parametrů v MA modelu 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 $N(0, \sigma^2), l = 0, ..., n, \phi_0 = 1$.

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

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

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 $X_i, i = 1, ..., n$. 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), které nabývá 1. až $(p+1)$. zaokrouhlená hodnota náhodného výběru.

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

$$p_{\mathbf{i}} = P(i_j - 0.5 \leq X_j < i_j + 0.5), \qquad j = 1, ..., p+1.$$

Nechť $A_{\mathbf{i}}$ je obdélník

$$A_{\mathbf{i}} = \Pi_{j=1}^{p+1}[i_j - 0.5; i_j + 0.5).$$

Jedná se o obdélník (od 3 rozměrů se jedná o kvádr) v prostoru, se středem v bodě $(\tilde{X}_1, ..., \tilde{X}_{p+1})$.

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 nabydou dvojice hodnot můžeme získat pomocí výpočtů nebo z geometrického zobrazení.

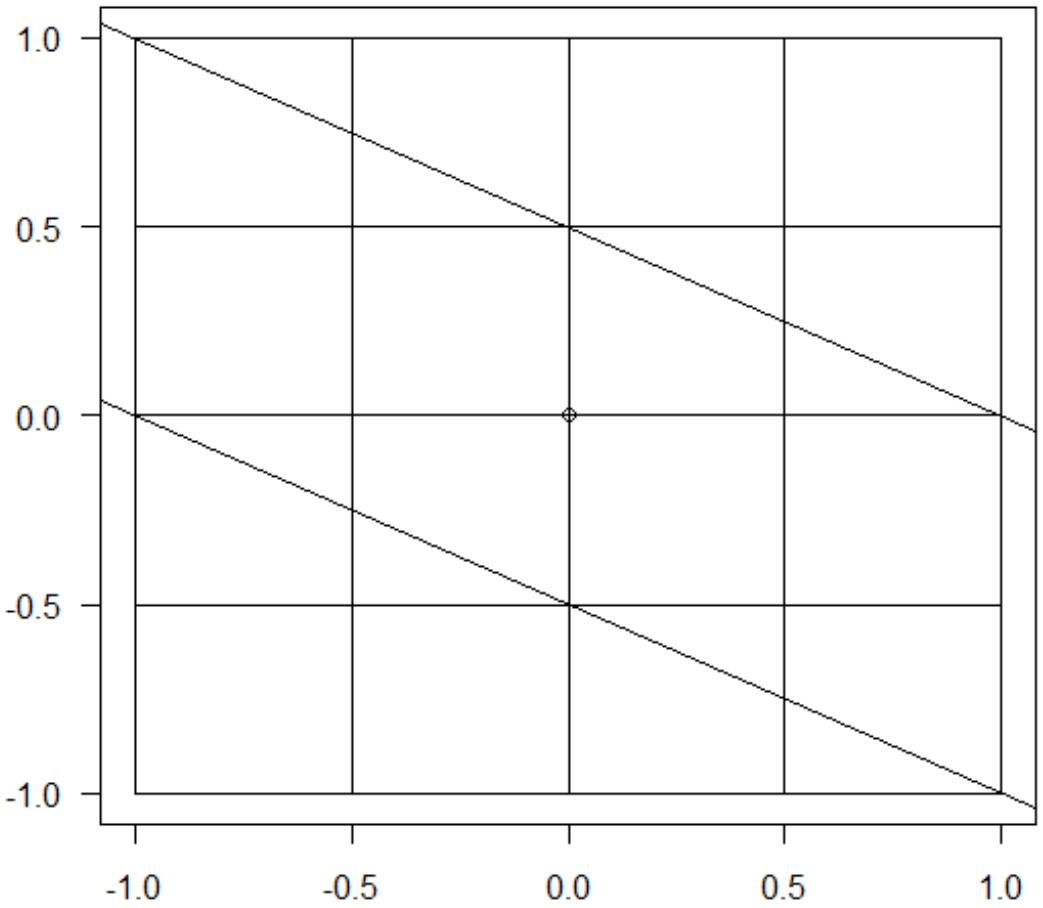
Nejprve získáme pravděpodobnosti první hodnoty na základě kombinací dvojic $\varepsilon_{t-1}, \varepsilon_t$. Pomocí grafického znázornění, kdy na osy naneseme hodnoty ε_{t-1} a ε_t dostaneme pravděpodobnosti první hodnoty (X_1) z dvojice (X_1, X_2). Jelikož ε_t pocházejí z rovnoměrného rozdělení, je pravděpodobnost nabytí hodnoty úměrná ploše v obdélníku $[-1, 1] \times [-1, 1]$. Pravděpodobnost nabytí hodnoty nula je polovina, pro hodnoty 1 a -1 je to čtvrtina.

Pokud má X_t hodnotu -1, pak ε_t musí ležet v intervalu $(-1, 0]$. V takovém případě pravděpodobnost nabytí druhé hodnoty z dvojice (X_{t+1}) určíme pouze na základě obdélníků nalevo od osy y — tyto hodnoty jsou popořadě $\frac{6}{16}, \frac{8}{16}, \frac{2}{16}$, pro -1,0 a 1. Pro X_t s hodnotou 1 musí ε_t ležet v intervalu $[0, 1)$. Pravděpodobnosti nabytí hodnoty X_{t+1} určíme díky symetrii rozdělení ε obdobně, jako pro ε_{t-1} o hodnotě -1. Pokud X_t má hodnotu 0, může ε_t ležet v celém intervalu $(-1, 1)$. Určení pravděpodobnosti nabytí hodnot X_{t+1} je stejné jako pro určení pravděpodobnosti nabytí hodnot 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 $\frac{6}{16}, \frac{8}{16}, \frac{2}{16}$, pro hodnotu 1 ze symetrie $\frac{2}{16}$, $\frac{8}{16}$ a $\frac{6}{16}$.

Obrázek 3.1: Grafické znázornění pravděpodobností pro kombinace $\varepsilon_i F$ (osa x) a ε_{i+1} (osa y) k příkladu MA(1) s ε z R$(-1, 1)$ modelu



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

$$\begin{aligned}
p_{(-1,1)} &= \frac{1}{4} \times \frac{6}{16} = \frac{3}{32}, \\
p_{(-1,0)} &= \frac{1}{4} \times \frac{8}{16} = \frac{4}{32}, \\
p_{(-1,1)} &= \frac{1}{4} \times \frac{2}{16} = \frac{1}{32}, \\
p_{(0,-1)} &= \frac{1}{2} \times \frac{1}{4} = \frac{4}{32}, \\
p_{(0,0)} &= \frac{1}{2} \times \frac{1}{2} = \frac{8}{32}, \\
p_{(0,1)} &= \frac{1}{2} \times \frac{1}{4} = \frac{4}{32}, \\
p_{(1,-1)} &= \frac{1}{4} \times \frac{2}{16} = \frac{1}{32}, \\
p_{(1,0)} &= \frac{1}{4} \times \frac{8}{16} = \frac{4}{32}, \\
p_{(1,1)} &= \frac{1}{4} \times \frac{6}{16} = \frac{3}{32}.
\end{aligned}$$

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 $(1,1)$.

Označme náhodný výběr $(X_1, ..., X_n)$ jako sekvenci $\boldsymbol{X}$ a náhodný výběr $(\tilde{X}_1, ..., \tilde{X}_n)$ jako sekvenci $\tilde{\boldsymbol{X}}$. Pak sekvence veličin $\boldsymbol{X}$ a $\tilde{\boldsymbol{X}}$ jsou p-závislé. Můžeme proto data rozdělit na $p+1$ 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) $\tilde{X}_{p+1}, ..., \tilde{X}_{2p+1},\ \tilde{X}_{3p+2}, ..., \tilde{X}_{4p+2}\ ...\ \tilde{X}_{m(p+1)}, ..., \tilde{X}_{m(p+1)+p}$.

Bai a kol. (2009) uvádějí postup konstruování odhadu ze zaokrouhlených dat $\tilde{X}_1, ..., \tilde{X}_n$ 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. Frekvence (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 $\sum_i n_i \ln p_i$. Tím získáme maximálně věrohodný odhad pravděpodobností jednotlivých dvojic hodnot p_i, z nich potom odvodíme maximálně věrohodné odhady parametrů. Výsledný maximálně věrohodný odhad označíme $(\hat{c_1}, \hat{\boldsymbol{\phi}}_{\mathbf{1}}, \hat{\sigma_1^2})$.

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 *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 $(p+1)$-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, $p+1$-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 5.1 ze skript Prášková (2001). 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. □

Věta 3.2 *Nechť*

$$\boldsymbol{\theta} = (c, \boldsymbol{\phi}, \sigma^2)^T,$$

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

$$I(\boldsymbol{\theta}) = \sum_{i=-\infty}^{\infty} [p_i(\boldsymbol{\theta})]^{-1} \frac{\mathrm{d}p_i(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \frac{\mathrm{d}p_i(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^T},$$

$$V(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) + 2 \sum_{t=2}^{p} \sum_{i_1, i_2} P(Y_1 \in A_{i_1}, Y_t \in A_{i_2}) p_{i_1}^{-1} p_{i_2}^{-1} \frac{\mathrm{d}p_{i_1}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \frac{\mathrm{d}p_{i_2}(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}^T},$$

$$G(\boldsymbol{\theta}) = I^{-1}(\boldsymbol{\theta}) V(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}).$$

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

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

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

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

Po zkonstruování odhadu průměrováním dílčích odhadů dostaneme

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

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 (ve výrazu $n-p$ zanedbáme p a uvažujeme pouze n). Dílčí odhady $\hat{\boldsymbol{\theta}_j}$ jsou konstruovány na základě $p+1$ nezávislých výběrů a mají stejné rozdělení jako $\hat{\boldsymbol{\theta}}$. Tedy

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

kde $\approx$ značí přibližnost. Označme

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

Úpravami dostaneme

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

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

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

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

Na pravé straně dostaneme

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

$m=\frac{n-p}{p+1}$, takže výrazy před sumou pod odmocninami dají přibližně n, hlavně pro $p<<n$ (zanedbáme p a místo $n-p$ počítáme s 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 $\sum_{j=1}^{n-p}\sum_{\boldsymbol{i}} I_{(Y_j\in A_{\boldsymbol{i}})}$, kde $I_{(Y_j\in A_{\boldsymbol{i}})}$ je indikátor, toho, že daná $(p+1)$-tice je v prostoru $A_{\boldsymbol{i}}$.

Pak můžeme použít vyjádření

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

Součty p-závislých sekvencí jsou normovány a můžeme použít centrální limitní větu. Z centrální limitní věty 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 $X_t = \varepsilon_t + \theta\varepsilon_{t-1}$, kde ε_t jsou nezávislé stejně rozdělené z N(0,σ^2), $|\theta| < 1$.

Z tvaru modelu

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

uvedeným Baiem a kol. (2009) ho dostaneme úpravou, jakou lze použít i u AR modelu - od náhodné veličiny X_t odečteme její střední hodnotu c. Tím dostaneme model bez aditivní konstanty, ve které má náhodná veličina nulovou střední hodnotu. Uvedenou úpravu je možno použít i pro MA model vyššího řádu než 1.

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

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

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

Guo a Li (2012) uvádějí, že věrohodnostní funce (θ, σ^2) na základě zaokrouhlených dat $\tilde{\boldsymbol{X}} = (\tilde{X}_1, ..., \tilde{X}_n)$ je definovaná vzorcem s hustotou a korigovaný MLE může být odvozen na základě $\mathbf{A} = \left| \begin{array}{cc} \frac{\partial^2 \ln f}{\partial\theta^2} & \frac{\partial^2 lnf}{\partial\theta\partial\sigma^2} \\ \frac{\partial^2 \ln f}{\partial\theta\partial\sigma^2} & \frac{\partial^2 lnf}{\partial(\sigma^2)^2} \end{array} \right|_{(\theta,\sigma^2)=(\tilde{\theta}_0,\tilde{\sigma^2}_0)}$,

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

kde

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

je pseudo MLE parametru

$$(\theta, \sigma^2)$$

. Pseudo maximálně věrohodným odhadem parametru θ nazývají odhad získaný řešením rovnice $\frac{\mathrm{d}lnf(x,\theta)}{\mathrm{d}\theta} = 0$, při dosazení zaokrouhlených dat, jako by šlo o přesná data.

Věta 3.3 *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

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

Výraz $\frac{-3\theta^2-1}{(\theta^2-1)^3} = \frac{3\theta^2+1}{(1-\theta^2)^3}$ nezávisí na t ani na n a proto ho lze přičíst k limitě sumy zbývající části výrazu.

Nyní budeme vyšetřovat výraz

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=2}^{n} \frac{(2t^2-t)\theta^{2t-2} + (-4t^2+6t+1)\theta^{2t} + (2t^2-5t+3)\theta^{2t+2}}{(\theta^2-1)^3}. \tag{3.1}$$

Postupným vyjádřením dostáváme

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

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 o součtu limit 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 $(n-1)$.

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

Pro n jdoucí k nekonečnu je výraz n^2 řádově vyšší než n nebo konstanta a $(n-1)/n$ konverguje k jedné. Tedy

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

Výraz n^2 konverguje pro rostoucí n k nekonečnu, výraz θ^{2n-2} pro $|\theta|<1$ jde k nule. Jelikož θ^{2n-2} konverguje rychleji než n^2, první sčítanec konverguje k nule. Stejným způsobem odvodíme i konvergenci dalších dvou sčítanců. Součet nul je definován, můžeme opět použít větu o limitě součtu. Protože jsme nalezli vyšší i nižší výraz než 3.1, který konverguje ke stejné hodnotě (0), podle věty o dvou strážnících je i limita výrazu 3.1 nulová.

Platí tedy

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

Věta 3.4 *Nechť*

$$e_t=\sum_{i=0}^{t-1}\theta^i y_{t-i},$$
$$\dot{e}_t=\sum_{i=0}^{t-1}i\theta^{i-1}y_{t-i},$$
$$\ddot{e}_t=\sum_{i=0}^{t-1}i(i-1)\theta^{i-2}y_{t-i}.$$

Potom máme:

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

kde

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

Důkaz:

$$\frac{1}{n}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t) = \frac{1}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2}y_{t-l}^2 + \frac{2}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-2} l(2l+1)\theta^{2l-1}y_{t-l}y_{t-l+1} + R_n,$$

kde

$$R_n = \frac{1}{n}\sum_{k=2}^{n-1}\sum_{t=k}^{n}\sum_{l=1}^{t-k} a_{l,t}(\theta)y_{t-l}y_{t-l+k}$$

s

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

pro všechna $2 \le k \le n-1$.

Z věty 3.3 a z Čebyševovy věty plyne

$$\frac{1}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-1} l(2l-1)\theta^{2l-2}y_{t-l}^2 \to_{n\to\infty} \frac{1+3\theta^2}{(1-\theta^2)^3}\gamma_0^* \tag{3.2}$$

v pravděpodobnosti.

Kde

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

kde U_i jsou zaokrouhlovací chyby.

Podobně

$$\frac{2}{n}\sum_{t=2}^{n}\sum_{l=1}^{t-2} l(2l+1)\theta_i^{2l-1}y_{i+1} \to_{n\to\infty} \frac{2\theta(3+\theta^2)}{(1+\theta^2)^3}\gamma_1^* \tag{3.3}$$

v pravděpodobnosti,

kde

$$\gamma_1 = E(\tilde{X}_t\tilde{X}_{t+1}) = E(X_tX_{t+1}) = -\theta_0\sigma_0^2.$$

Konečně, když

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

a z Čebyševovy věty plyne

$$R_n \to_{n\to\infty} 0 \tag{3.4}$$

v pravděpodobnosti. Výsledek 1) plyne z 3.2, 3.3 a 3.4. Podobně dostaneme výsledek 2). □

Věta 3.5 *Nechť*

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

je PMLE (θ, σ^2) *pro MA(1) model pak*

$$1)\ \lim_{n\to\infty}\frac{1}{n}\boldsymbol{A} = -\frac{1}{\hat{\sigma}_0^2}\,\mathrm{diag}\left(\frac{(1+3\hat{\theta}_0^2)(1+\hat{\theta}_0^2)\hat{\sigma}_0^2 + 2\hat{\theta}_0(3+\hat{\theta}_0^2)(-1+\hat{\theta}_0\hat{\sigma}_0^2)}{(1-\hat{\theta}_0^2)^3}, \frac{1}{2\hat{\sigma}_0^2}\right),$$

$$2)\ \lim_{n\to\infty}\frac{1}{n}\boldsymbol{b}^T = \frac{h^2}{12\hat{\sigma}_0^4}\left(\frac{\hat{\theta}_0}{(1-\hat{\theta}_0^2)^2}, -\frac{1}{2(1-\hat{\theta}_0^2)}\right).$$

Důkaz: 1) Z normální rovnice máme

$$-\frac{1}{\sigma^2}\sum_{t=2}^{n} e_t\dot{e}_t\Big|_{\hat{\beta}_0} = 0,$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{t=1}^{n} e_t^2\Big|_{\hat{\beta}_0} = 0$$

Z druhé parciální derivace $\ln f$ vzhledem k θ a σ^2 jsou

$$\frac{\partial^2 \ln f}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{t=1}^{n} e_t^2 = -\frac{n}{2\hat{\sigma}_0^4},$$

$$\frac{\partial^2 \ln f}{\partial\theta\partial\sigma^2} = \frac{1}{\sigma^4}\sum_{t=2}^{n}(\dot{e}_t^2 + e_t\ddot{e}_t).$$

Z věty 3.4 a tvaru matice $\boldsymbol{A}$ plyne výsledek 1).
2) Poznamenejme že:

$$\frac{\partial}{\partial\theta}\frac{\partial^2 ln(f)}{\partial y_t^2} = -\frac{2\theta^{2n-2t+1}[\theta^{-2(n-t)+(n-t)\theta^2-(n-t-1)}]}{\sigma^2(1-\theta^2)^2},$$

$$\frac{\partial}{\partial\sigma^2}\frac{\partial^2 ln(f)}{\partial y_t^2} = \frac{1}{\sigma^4}\sum_{i=0}^{n-t}\theta^{2i},$$

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

$$\frac{\partial}{\partial\sigma^2}\left(\frac{\partial ln(f)}{\partial y_t}\right)^2 = -\frac{2}{\sigma^6}\sum_{i=0}^{n-t}\theta^i e_{t+i}.$$

A z aproximace b_j v korigovaném MLE plyne

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

$$b_2 = -\frac{h^2}{24\hat{\sigma}_0^4}\frac{\hat{\theta}_0^{2+2n} - \hat{\theta}_0^2 + (1-\hat{\theta}_0^2)n}{(1-\hat{\theta}_0^2)^2}.$$

Pak limity $\frac{b_1}{n}$ a $\frac{b_2}{n}$ mohou být . □
Konečně, z věty 3.5 a definice korigovaného MLE plyne následující věta:

Věta 3.6 *Korigovaný MLE modelu MA(1) založený na zaokrouhlených datech je:*

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

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

kde $\hat{\theta}_0, \hat{\sigma}_0^2$ 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) ve větě 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)).

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}.

Když $|\phi| < 1$, můžeme najít dostatečně velké k, aby $|\phi|^k$ bylo tak malé, jak potřebujeme. 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:

1) $(\tilde{X}_1, \tilde{X}_2)$, $(\tilde{X}_2, \tilde{X}_3)$,...,$(\tilde{X}_{(m-1)k+1}, \tilde{X}_{(m-1)k+2})$

2) $(\tilde{X}_2, \tilde{X}_3)$, $(\tilde{X}_3, \tilde{X}_4)$,...,$(\tilde{X}_{(m-1)k+2}, \tilde{X}_{(m-1)k+3})$

... .

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 $(\tilde{X}_1, \tilde{X}_2)$,$(\tilde{X}_2, \tilde{X}_3)$,...,$(\tilde{X}_{(m-1)k+1}, \tilde{X}_{(m-1)k+2})$ 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 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:

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

Ke konstrukci AMLE rozvedeme krok 1). Odhady získáme na základě řešení soustavy rovnic

$$\sum_i \sum_j \frac{n_{ij}^{(1)}}{p_{ij}} \frac{\partial p_{ij}}{\partial c} = 0,$$
$$\sum_i \sum_j \frac{n_{ij}^{(1)}}{p_{ij}} \frac{\partial p_{ij}}{\partial \phi} = 0,$$
$$\sum_i \sum_j \frac{n_{ij}^{(1)}}{p_{ij}} \frac{\partial p_{ij}}{\partial \sigma^2} = 0.$$

Věta 3.7 *Pokud zaokrouhlíme veličiny generované AR(1) modelem, budou se chovat jako nezaokrouhlené veličiny v modelu ARMA(1,1).*

Vyjděme z AR(1) modelu ve tvaru $X_t = \phi X_{t-1} + \varepsilon_t$, kde $X_i, i \in N$ mají nulovou střední hodnotu. Na tento tvar lze AR(1) model převést z původního tvaru

$$X_t^* = c + \phi_1 X_{t-1}^* + \varepsilon_t$$

tím, že od obou stran odečteme střední hodnotu veličiny X^*, tedy $\frac{c}{1-\phi}$.

Tedy

$$X_t^* - \frac{c}{1-\phi} = c + \phi_1\left(X_{t-1}^* - \frac{c}{1-\phi}\right) + \phi\frac{c}{1-\phi} - \frac{c}{1-\phi} + \varepsilon_t.$$

Úpravou pravé strany dostaneme

$$\begin{aligned} &c + \phi_1(X_{t-1}^* - \phi_1 X_{t-1}^*) + \phi\frac{c}{1-\phi} - \frac{c}{1-\phi} + \varepsilon_t \\ &= \varepsilon_t + \phi_1(X_{t-1}^* - \frac{c}{1-\phi}) + c(1 + \frac{\phi}{1-\phi} - \frac{1}{1-\phi}) \end{aligned}$$

Jelikož $1 + \frac{\phi}{1-\phi} - \frac{1}{1-\phi} = 0$, po přeznačení $X_i^* - \frac{c}{1-\phi} = X_i$ dostaneme AR(1) model s veličinami o nulové střední hodnotě a s nulovou aditivní konstantou.

Uvažujme model ARMA(1,1) ve tvaru

$$Y_t = b + \epsilon_t + \delta_1 Y_{t-1} + \theta_1\epsilon_{t-1}.$$

Model ARMA také můžeme (podobně jako AR) převést na tvar s nulovou aditivní konstantou. Z ARMA modelu vyjádříme střední hodnotu veličiny Y: $EY = b + E\epsilon + E\delta_1 Y + E\theta_1\epsilon$ $EY = b + \delta_1 EY$ $EY = \frac{b}{1-\delta_1}$

Ve vyjádření

$$Y_t = b + \epsilon_t + \delta_1 Y_{t-1} + \theta_1\epsilon_{t-1}.$$

odečteme od obou stran $\frac{b}{1-\delta_1}$ a výraz doupravíme, abychom na osbou stranách místo veličiny Y dostali její centrovanou verzi (s nulovou střední hodnotou).

$$\begin{aligned} Y_t - \frac{b}{1-\delta_1} &= b + \epsilon_t + \delta_1\left(Y_{t-1} - \frac{b}{1-\delta_1}\right) + \delta_1\frac{b}{1-\delta_1} + \theta_1\epsilon_{t-1} - \frac{b}{1-\delta_1} \\ Y_t - \frac{b}{1-\delta_1} &= \left(\frac{b(1-\delta_1) + b\delta_1 - b}{1-\delta_1}\right) + \epsilon_t + \delta_1\left(Y_{t-1} - \frac{b}{1-\delta_1}\right) + \theta_1\epsilon_{t-1} \\ Y_t - \frac{b}{1-\delta_1} &= \epsilon_t + \delta_1\left(Y_{t-1} - \frac{b}{1-\delta_1}\right) + \theta_1\epsilon_{t-1}. \end{aligned}$$

Nyní dokážeme, že zaokrouhlená data z AR(1) modelu se chovají jako nezaokrouhlená data z 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 obecnějším tvaru $\sum_{i=0}^{p} a_i X_{t-i} = \sum_{j=0}^{m} b_j \varepsilon_{t-j}$. a spektrální hustotu ve tvaru $f(\lambda) = \frac{\sigma^2}{2\pi}\frac{|b_0+b_1e^{-i\lambda}+\ldots+b_me^{-im\lambda}|^2}{|a_0+a_1e^{-i\lambda}+\ldots+a_ne^{-in\lambda}|^2}$ než v předchozím uvedeném, pro který platí $a_0 = b_0$.

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

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

Aby pro upravený tvar platilo $f_z(\lambda) = f_x(\lambda) + f_y(\lambda)$ musí platit

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

Vydělením rovnic a po úpravě dostaneme $b = \frac{D(a-b)(1-ab)}{2\pi A}$. Upravíme a vyřešíme uvedenou 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 ($A = \frac{C(1-a^2)}{2\pi}$).

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

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 z R je kladná a z omezení na a a kladnosti C a D je kladný i druhý člen, 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, musí být jeden z nich v absolutní hodnotě menší než 1.

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

Protože součin kořenů je 1, musí mít oba kořeny stejné znaménko (oba jsou buď záporné, anebo kladné).

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 $2Da$. 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 ukázali předtím. Jeho odmocnina je také kladné číslo a součet samých kladných čísel je kladný. Kořeny tedy mají stejné znaménko jako a. □

Simulovali 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. 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 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_1 hodnotu $0,509896$, pro θ_1 hodnotu $-0,041563$ a pro konstantu b hodnotu 0.002915. Při vyšetření nezaokrouhlených dat jsme získali výsledky $0,512525$, $0,006958$ a $-0,004834$.

Při hledání vhodného modelu pomocí Akaikeho kritéria pro model ARMA(p,q), s hodnotami $p, q \leq 5$, vyšla jak pro nezaokrouhlená, tak pro zaokrouhlená data nejnižší hodnota AIC pro model AR(1). Pro nezaokrouhlená data mělo Akaikovo kritérium hodnotu 2838,653 , pro zaokrouhlená data 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_1 hodnotu $0,5093$, pro θ_1 hodnotu $-0,0414$ a pro konstantu b hodnotu $0,0061$. Při vyšetření nezaokrouhlených dat jsme získali výsledky $0,5116$, $0,0076$ a $-0,0112$.

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 kritéri 116,4.

Při vyšetřování pouze mezi AR modely do řádu 5 pomocí Yule-Walkerovy metody dostaneme v obou případech model AR(1). Pro nezaokrouhlená data s hodnotou koeficientu 0,517 a hodnotou Akaikeho kritéria 0, pro zaokrouhlená data s hodnotou koeficientu 0,4775 a taktéž s nulovou hodnotou Akaikeho kritéria.

Při vyšetřování metodou maximální věrohodnosti do řádu 5 funkcí `ar` vyjde v obou případech také jako nejlepší model AR(1). Vypočítaná hodnota koeficientu ϕ_1 je 0,5171 pro nezaokrouhlená data a 0,4774 pro zaokrouhlená data. Hodnota AIC je pro AR(1) v obou případech nulová.

Při odhadu koeficientu AR(1) modelu při známé střední hodnotě pomocí funkce `AR1Est` vyšlo 0,51718 pro nezaokrouhlená a 0,4774 pro zaokrouhlená data.

Při vyšetřování statistik modelu AR(1) pomocí metody maximální věrohodnosti získáme pro nezaokrouhlená data logaritmickou věrohodnost 2,394 a AIC $-0,8$, pro zaokrouhlená data logaritmickou věrohodnost $-55,355$ a AIC 114,7. U nezaokrouhlených dat vychází menší střední čtvercová chyba, ale střední hodnota modelu ze zaokrouhlených dat je bližší skutečné hodnotě (0,007 oproti $-0,0094$, střední hodnota modelu je nulová).

Při použití věrohodnostní metody McLeoda a Zhanga (2007) dostaneme logaritmickou věrohodnost 2,39 pro nezaokrouhlená data a $-55,36$ pro zaokrouhlená data. Hodnota Akaikeho kritéria je $-0,8$ pro nezaokrouhlená a 114,7 pro zaokrouhlená data.

Při odhadu koeficientů v modelu ARMA(1,1) vychází odhad koeficientu θ u nezaokrouhlených dat řádově vyšší . Přesto vychází v obou případech (pro $p, q \leq 5$) nejnižší Akaikeho kritérium pro model AR(1).

Při vyšetřování mezi AR modely je model (vyjma použití metody maximální věrohodnosti) v obou případech správně identifikován jako AR(1). S výjimkou dvou případů, kdy Akaikeho kritérum vyšlo pro oboje data nulová, byla hodnota Akaikeho kritéria pro nezaokrouhlená data menší, v některých případech dokonce i řádově.

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

Anděl (1976) dokazuje, že data s nezávislým šumem se tak chovají. 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 .

3.2.1 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 3.1 a 3.2 . 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 $\beta = (\mu, \boldsymbol{\phi}, \sigma^2), \boldsymbol{\phi} = (\phi_1, ..., \phi_p)$. 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 $[-h/2; h/2]$.

Guo a Li (2012) uvádějí jiné varianty odhadů vycházejících z maximálně věrohodného odhadu — pseudo maximálně věrohodný odhad (PMLE) a korigovaný maximálně věrohodný odhad. 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 $L(\tilde{\boldsymbol{x}}, \theta) = h^{-n} \int_{\tilde{x}_n - h/2}^{\tilde{x}_n + h/2} ... \int_{\tilde{x}_1 - h/2}^{\tilde{x}_1 + h/2} f(\boldsymbol{u}, \theta) \, \mathrm{d}u_1 ... \, \mathrm{d}u_n$.

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

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

$L(\beta, y) = f(y, \beta) + h^2/24 \sum_{t=1}^{n} \frac{\partial^2 f(y,\beta)}{\partial y_t^2} + O(h^3)$. Guo a Li (2012) uvádějí, že logaritmickou věrohodnost lze aproximovat v blízkosti $h = 0$ pomocí

$$L(\beta, y) \sim \ln f(y, \beta) + h^2/24 \sum_{t=1}^{n} [\frac{\frac{\partial^2 f(\boldsymbol{y},\beta)}{\partial y_t^2}}{f(\boldsymbol{y}, \beta)}] + O(h^3)$$

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

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

kde

$$\begin{aligned} A &= [a_{ij}], \boldsymbol{b} = [b_j], \\ a_{ij} &= \frac{\partial^2 \ln(L(\boldsymbol{y}, \theta))}{\partial\theta_i \partial\theta_j} \approx \left.\frac{\partial^2 \ln(f(\boldsymbol{y}, \theta))}{\partial\theta_i \partial\theta_j}\right|_{\theta=\hat{\theta}_0}, \\ b_j &= \frac{\partial ln(L(\boldsymbol{y}, \theta))}{\partial\theta_j} \approx \frac{h^2}{24} \frac{\partial}{\partial\theta_j} \left.\sum_{t=1}^{n} [\frac{\frac{\partial^2 f(\boldsymbol{y},\theta)}{\partial y_t^2}}{f(\boldsymbol{y}, \theta)}]\right|_{\theta=\hat{\theta}_0}, \end{aligned}$$

kde θ_i je i-tá složka θ, $i, j = 1, ..., p + 2$.

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 rozdělné z N(0,σ^2). Odhadovaný parametr je $\boldsymbol{\theta} = (\mu, \phi_1, ..., \phi_p, \sigma^2)^T$. Stam a Cogger (1993) uvádějí věrohodnostní funkce pro přesná data. Její tvar je

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

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

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

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

$$L_1(\boldsymbol{\theta}, \tilde{\boldsymbol{X}}) = \int_{\tilde{x}_n - h/2}^{\tilde{x}_n + h/2} ... \int_{\tilde{x}_1 - h/2}^{\tilde{x}_1 + h/2} f(\boldsymbol{X}, \boldsymbol{\theta}) dX_1 ... dX_n,$$

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

Pro j-tý prvek parametru získáme odhad

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

Stam a Cogger (1993) uvádějí, že maximálně věrohodný odhad $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\boldsymbol{X})$ nelze vypočítat, protože nejsou známa přesná data. Po vzoru Lindleyho (1950) odvozují $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\tilde{\boldsymbol{X}})$. Odhady $\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\tilde{\boldsymbol{X}})$ jsou brány jako výchozí bod pro použití Newton-Rawsonovy 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

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

a $\boldsymbol{b}$ vektor s prvky

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

se známými hodnotami v $\hat{\boldsymbol{\theta}}_{\boldsymbol{0}}(\tilde{\boldsymbol{X}})$.

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

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

Úpravou a zlogaritmováním dostaneme

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

Pro malé x je $\ln(1 + x) \approx ln(x)$ a po zderivování předchozího vzorce podle θ_i dostaneme rovnici

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

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

$$a_{ij} \doteq \frac{\partial^2 \ln(f(\tilde{\boldsymbol{X}}, \boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j}\bigg|_{\boldsymbol{\theta}}$$
$$= \hat{\boldsymbol{\theta}}(\tilde{\boldsymbol{X}}), \qquad i, j = 1, ..., p + 2$$

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

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.

Nejprve budeme uvažovat případ, kdy $n = 100$, $\mu = 0$, $\sigma^2 = 25$, $\mu_\varepsilon = 0$, $\beta_0 = 4, \beta_1 = 3$.

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 .

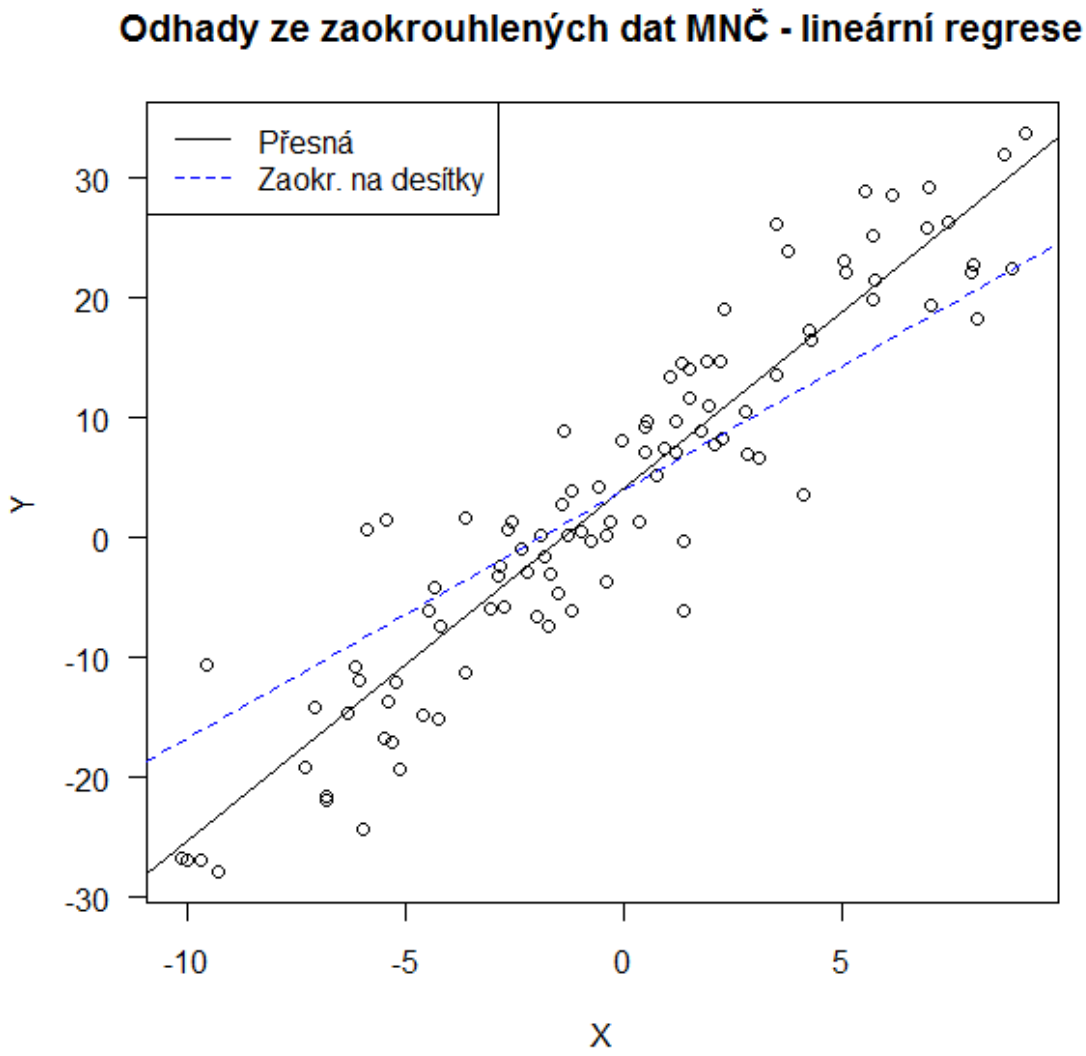
Po vykreslení přímek je na grafu 4.1 vidět, že od přímky s parametry odhadnutými z nezaokrouhlených pozorování je rozlišitelná pouze přímka s parametry odhadnutými z pozorování zaokrouhlenými na desítky.

Pokud pracujeme se zaokrouhlenými daty, je vhodnější namísto metody nejmenších čtverců odhadnout parametry přímky pomocí momentové metody

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

Obrázek 4.1: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



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 $y_i = \alpha + \beta x_i, i = 1, ..., n$. Namísto dvojic (x_i, y_i) můžeme pozorovat pouze dvojice (ξ_i, η_i), které jsou zatíženy náhodnými chybami.

Budeme předpokládat, že $\xi_i = x_i + \delta_i$, $\eta = y_i + \gamma_i$, $\quad i = 1, ..., n$, kde $\delta_i, i = 1, ..., n$ pochází z rovnoměrného rozdělení $R(a, b)$ a $\gamma_i, i = 1, ..., n$ pochází z normálního rozdělení $N(0, \sigma_\gamma^2)$.

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:

$$\eta_i = \alpha\beta\xi_i + (\gamma_i - \beta\delta_i), i = 1, ..., n.$$

Jelikož

$$\operatorname{cov}(\xi_i, \gamma_i - \beta\delta_i) = \operatorname{cov}(x_i + \delta_i, \gamma_i - \beta\delta_i) = -\beta\sigma_\delta^2$$

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

$$\begin{aligned}
E\xi_i &= E(x_i + \delta_i) = Ex_i + E\delta_i = \mu + 0 = \mu, \\
E\eta_i &= E(y_i + \gamma_i) = E\eta = y_i + E\gamma_i = \alpha + \beta\mu + 0 = \alpha + \beta\mu \\
\operatorname{cov}(\xi_i, \eta_i) &= \operatorname{cov}(x_i + \delta_i, y_i + \gamma_i) = \operatorname{cov}(x_i + \delta_i, \alpha + \beta x_i + \gamma_i) \\
&= \operatorname{cov}(x_i + \delta_i, \beta x_i + \gamma_i) = \beta\sigma_x^2, \\
\operatorname{var}(\xi_i) &= \operatorname{var}(x_i + \delta_i) = \operatorname{cov}(x_i + \delta_i, x_i + \delta_i) = \sigma_x^2 + \sigma_\delta^2 \\
\operatorname{var}(\eta_i) &= \operatorname{var}(y_i + \gamma_i) = \operatorname{var}(\alpha + \beta x_i + \varepsilon_i) \\
&= \operatorname{cov}(\alpha + \beta x_i + \varepsilon_i, \alpha + \beta x_i + \varepsilon_i) = \beta^2\sigma_x^2 + \sigma_\varepsilon^2.
\end{aligned}$$

Odhady parametrů získáme momentovou metodou.

Dostáváme:

$$\begin{aligned}
\mu &= \bar{\xi}, \\
\alpha + \beta\mu &= \bar{\eta}, \\
\beta\sigma_x^2 &= s_{\xi\eta}, \\
\sigma_x^2 + \sigma_\delta^2 &= s_\xi^2, \\
\beta^2\sigma_x^2 + \sigma_\gamma^2 &= s_\eta^2,
\end{aligned}$$

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

kde $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$ je průměr, $s_{\xi,\eta} = \frac{1}{n}\sum_{i=1}^{n}\left((\xi_i - \bar{\xi})(\eta_i - \bar{\eta})\right)$ je výběrová kovariance a $s_\xi^2 = s_{\xi,\xi}$ 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 $(-5, 5)$.

Pomocí Kolmogorovova-Smirnovova testu otestujeme hypotézu, že $\boldsymbol{X} - \tilde{\boldsymbol{X}}$ (pro všechny stupně jemnosti zaokrouhlení použité v tabulce 4.1) pochází z rovnoměrného rozdělení o délce odpovídající vzdálenosti dvou nejbližších různých diskrétních hodnot, na které zaokrouhlujeme, se středem v nule, proti oboustranné alternativě. (Tedy pro zaokrouhlení na desítky testujeme H_0, že $\boldsymbol{X} - \tilde{\boldsymbol{X}}$ pochází z R(-5 ; 5), pro zaokrouhlení na jednotky z R(-0.5 ; 0.5), atd.)

Výsledky testu nejsou na hladině $\alpha = 0.05$ v rozporu s předpokladem, že chyby zaokrouhlování pocházejí z rovnoměrného rozdělení.

Z očekávaného rozdělení δ 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 $\sigma_\delta^2 = \frac{100}{12} = 8,333...$.

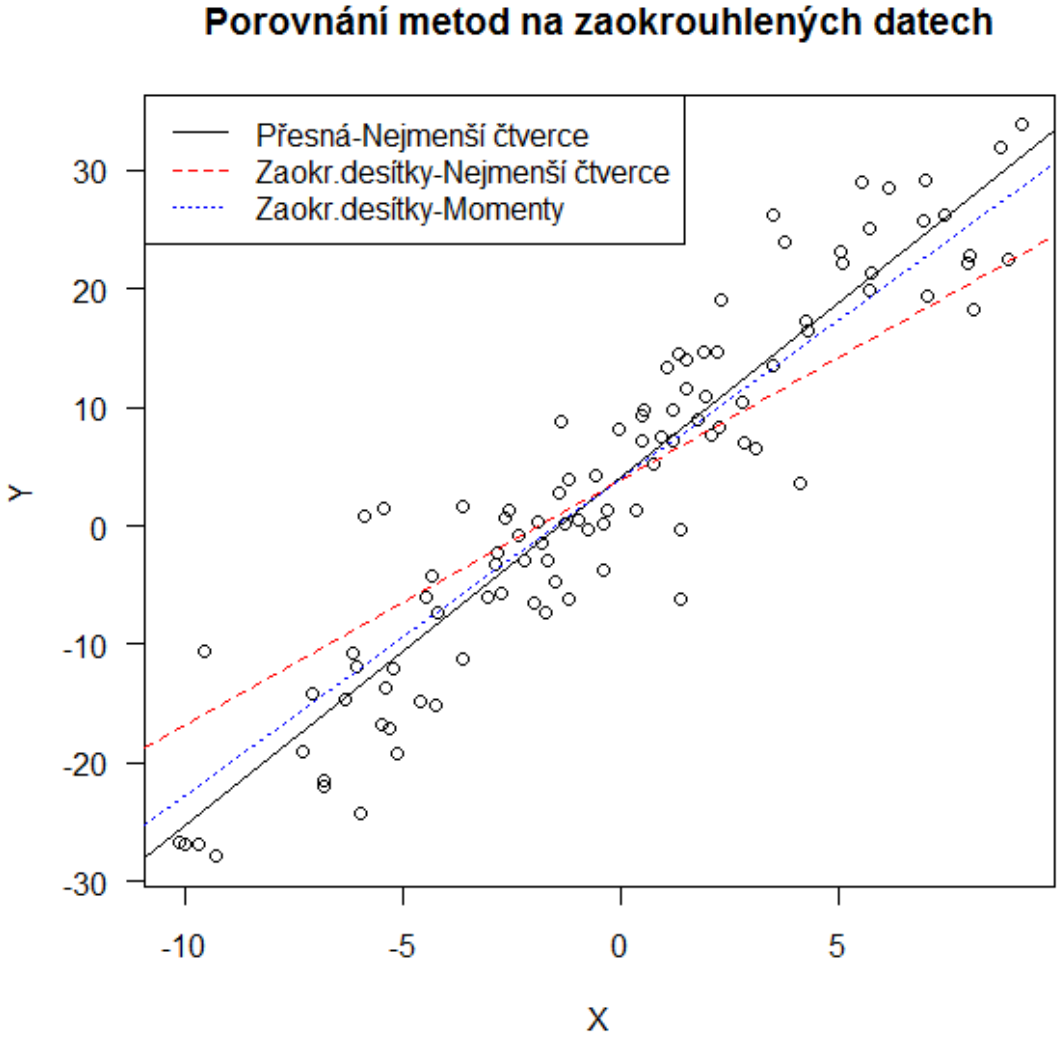
Tím dostáváme pět rovnic pro pět neznámých. Jejich vyřešením z náhodného výběru $x_1, ..., x_{100}$ zaokrouhleného na desítky (tedy $\xi_1, ..., \xi_{100}$) dostaneme odhady $\hat{\beta_0} = 4.032524, \hat{\beta_1} = 2.675403$.

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.

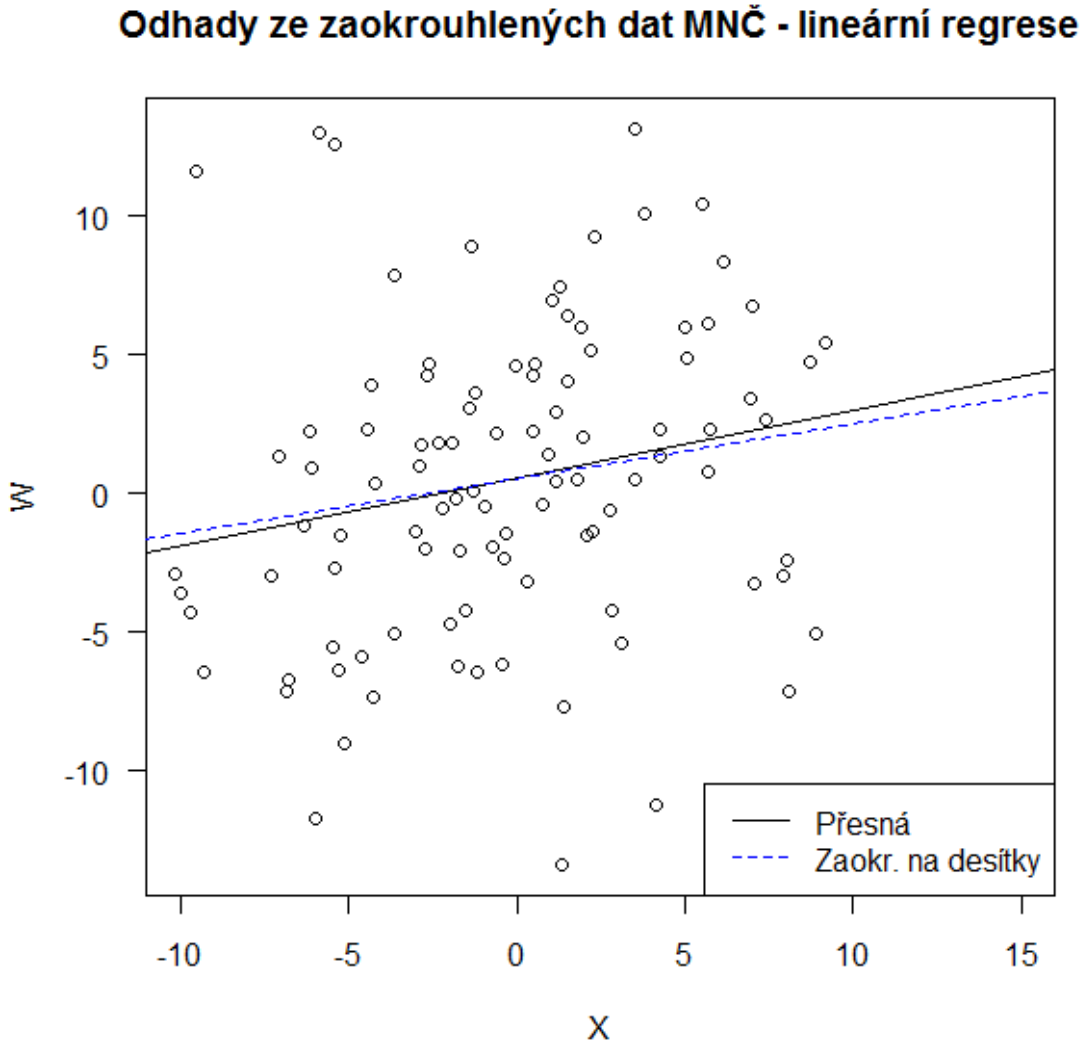
Obrázek 4.2: Porovnání metody nejmenších čtverců a momentové metody pro zaokrouhlená data



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

Obrázek 4.3: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



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

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

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 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.

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

Tentokrát na hladině $\alpha = 0.05$ zamítneme nulovou hypotézu u dat zaokrouhlených na desetiny.

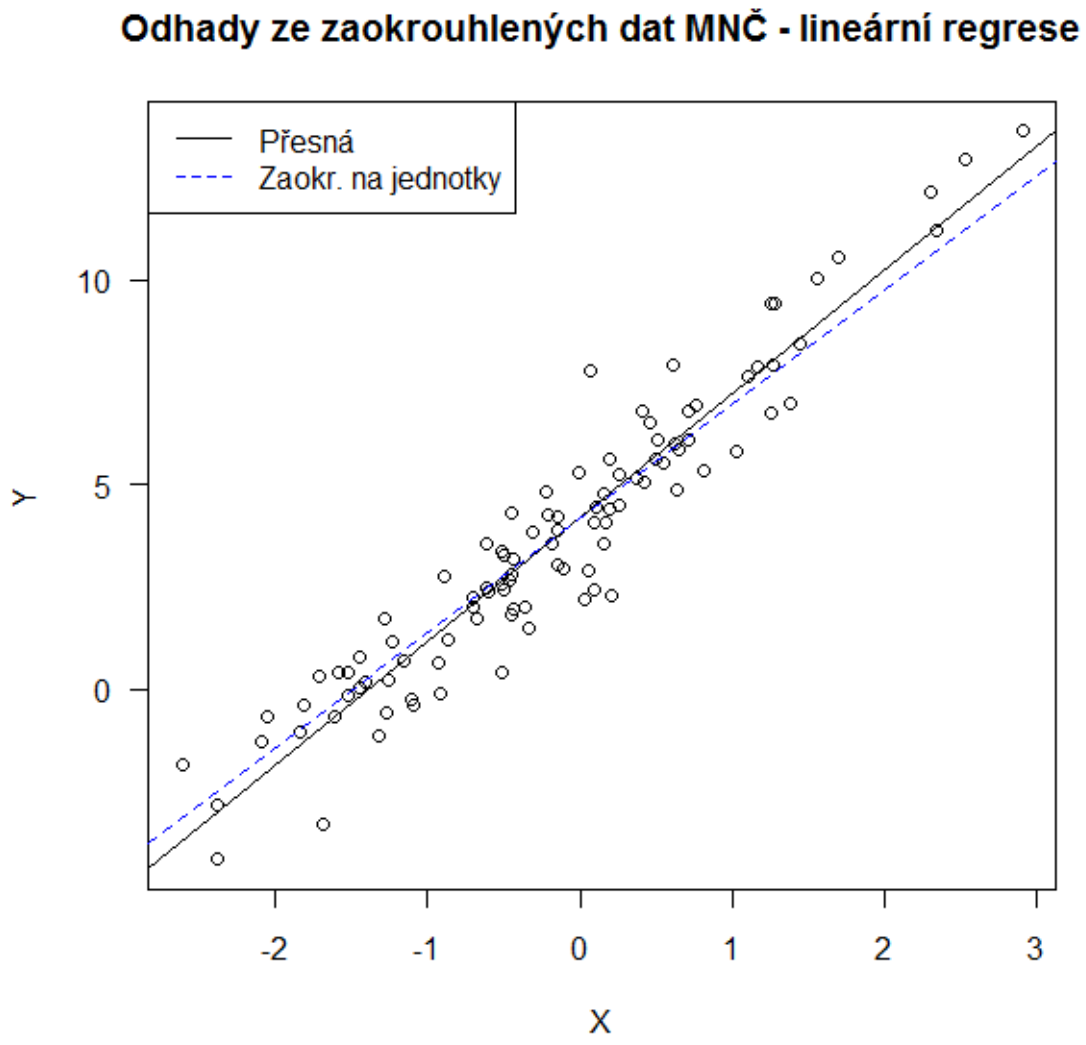
Jelikož u zaokrouhlení na jednotky jsme nulovou hypotézu nezamítli, budeme dále pro použití metody momentů předpokládat, že chyba δ pochází z $R(-0.5, 0.5)$, z čehož dostaneme její rozptyl $\sigma_\delta^2 = \frac{1}{12} = 0.08333...$.

Dostáváme odhady parametrů $\hat{\beta_0} = 4.211866$, $\hat{\beta_1} = 2.983287$.

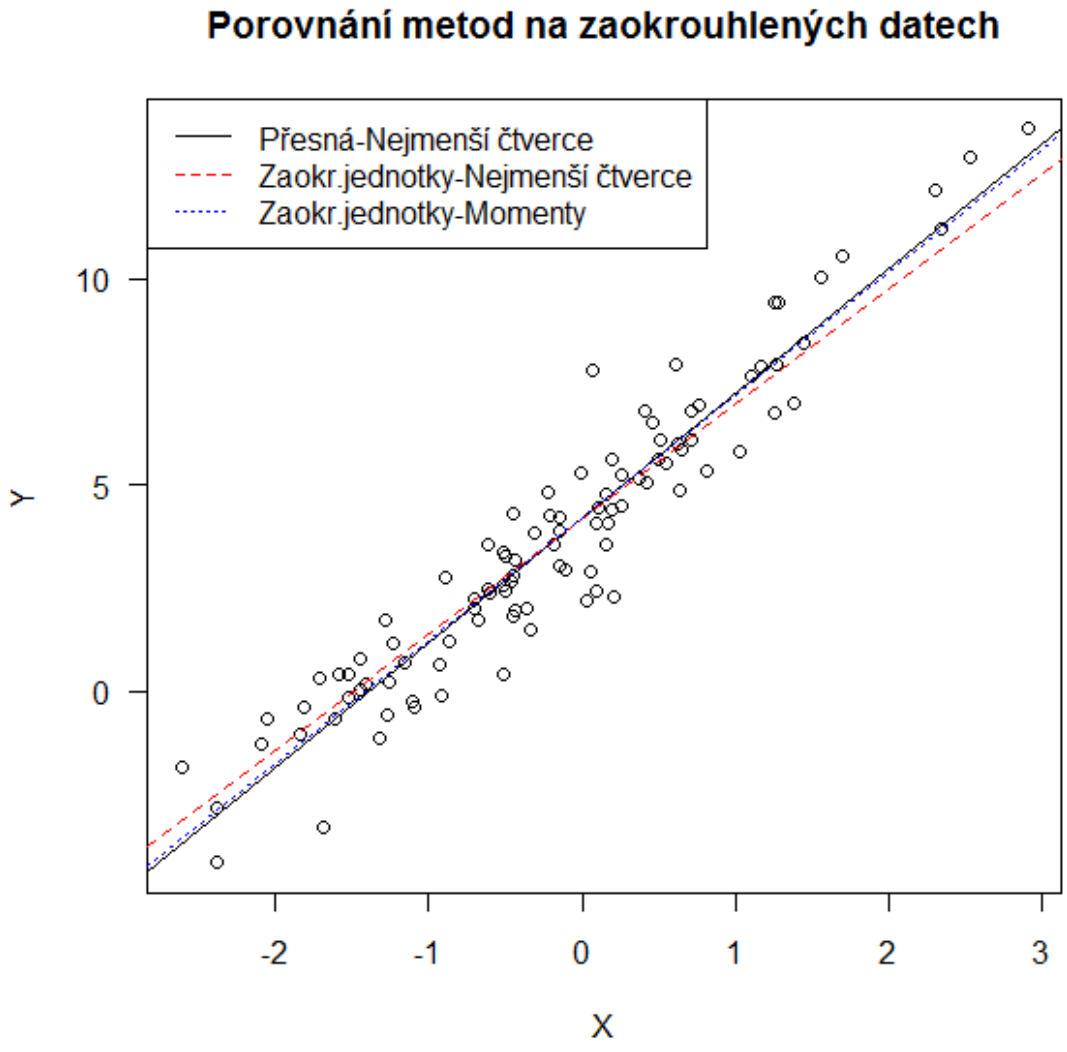
Na obrázku 4.5 je opět vidět, že momentová metoda dává lepší výsledek než metoda nejmenších čtverců.

Uvažujme opět dvojice veličin (x_i, y_i), mezi nimiž platí lineární vztah $y_i = \alpha + \beta x_i, i = 1, ..., n$. K dispozici máme opět pouze dvojice (ξ_i, η_i) zatížené náhodnými chybami. Dále předpokládáme, že platí $\xi_i = x_i + \delta_i$, $\eta = y_i + \varepsilon_i, i = 1, ..., n$.

Obrázek 4.4: Přímky s parametry odhadnutými MNČ ze zaokrouhlených dat



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



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

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 shrnuty 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 na námi zvolených datech vykazuje nedostatky o trochu menší než metoda

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

nejmenších čtverců.

Literatura

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

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

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

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

[5] 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.

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

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

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

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

[10] Prášková, Z. (2001): Základy náhodných procesů II. Karolinum, Praha.

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

[12] 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.

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

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

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

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

[17] www.wolframalpha.com