Procjena autokorelacijske funkcije

Paprić, Doris

Master's thesis / Diplomski rad

2020

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **Josip Juraj Strossmayer University of Osijek, Department of Mathematics / Sveučilište Josipa Jurja Strossmayera u Osijeku, Odjel za matematiku**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:126:769181

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-28**



Repository / Repozitorij:

Repository of School of Applied Mathematics and Computer Science







Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Doris Paprić

Procjena autokorelacijske funkcije

Diplomski rad

Osijek,2020.

Sveučilište J.J. Strossmayera u Osijeku
Odjel za matematiku

Doris Paprić

Procjena autokorelacijske funkcije

Diplomski rad

Mentor: prof.dr.sc. Mirta Benšić

Komentor: dr.sc. Ivan Papić

Osijek,2020.

Sadržaj

Uvod **2**

1 Vremenski nizovi **3**

1.1 Stacionarni procesi 3

1.2 Modeli stacionarnih vremenskih nizova 6

2 Procjena očekivanja **9**

3 Uzoračka funkcija autokorelacija **11**

3.1 Primjeri 14

4 Robusni procjenitelji autokorelacijske funkcije **18**

4.1 L1 procjenitelj 22

4.2 Procjenitelj metodom najmanjih odrezanih kvadrata 23

4.3 MM procjenitelj 25

5 Usporedba procjenitelja na primjerima **28**

Literatura **35**

Sažetak **36**

Summary **37**

Životopis **38**

Uvod

Promatrajući neku pojavu tijekom vremena često možemo naslutiti postojanje neke vrste veze. Za razliku od korelacije, koja je jedna od matematičkih procedura za objašnjavanje povezanosti između dvije slučajne varijable, autokorelacija nam govori o korelaciji slučajne varijable sa zaostalim vrijednostima same sebe. Najčešće se raspravlja u kontekstu vremenskih nizova kao nizova podataka promatranih u uzastopnim vremenskim trenucima. Primjena autokorelacije je široka, koristi se u medicini, astrofizici, snimanju glazbe, GPS sustavu... Vrlo česta primjena je u financijskoj matematici. Na primjer, ako investitori znaju da dionica ima povijesno visoku pozitivnu vrijednost autokorelacije i svjedoče da je ona postigla znatne dobitke u posljednjih nekoliko dana, tada mogu s razlogom očekivati da će se kretanje tijekom narednih nekoliko dana podudarati sa zaostalim vrijednostima vremenskog niza i tako ostvariti zaradu.
Kada raspolažemo s vremenskim nizom, funkcija autokorelacija nam je nepoznata, no možemo odrediti njezinog procjenitelja, odnosno uzoračku funkciju autokorelacija. Kako su mnogi slučajni procesi karakterizirani funkcijom očekivanja, autokovarijanci i autokorelacija, procjenitelji tih funkcija imaju važnu ulogu pri zaključivanju o procesu, a posebno u problemu konstruiranja odgovarajućeg modela za podatke. Cilj ovog diplomskog rada je opisati procjenu autokorelacijske funkcije te navesti nekoliko vrsta procjenitelja i njihova svojstva.
U prvom dijelu rada objašnjena je teorijska osnova o vremenskim nizova i koncept stacionarnosti. Kako funkcija očekivanja ima važnu ulogu u definiranju procjenitelja autokorelacija, u drugom poglavlju napravljena je procjena očekivanja.
U glavnom dijelu navedeni su procjenitelji autokorelacijske funkcije. Najprije je opisan klasični procjenitelj i njegova teorijska svojstva, a primjena istih prikazana je na primjerima najznačajnijih procesa. Osim klasičnog procjenitelja, uvedeni su i robusni procjenitelji temeljeni na robusnim regresijskim tehnikama. Promatrani su L1 procjenitelj, procjenitelj metodom najmanjih odrezanih kvadrata i MM procjenitelj.
Zadnji dio se odnosi na ilustraciju teoretskih rezultata kroz simulacije nekih stacionarnih procesa.

1 Vremenski nizovi

Promatrati neku pojavu tijekom vremena znači bilježiti njezine vrijednosti u određenim vremenskim trenucima i analizirati promjene koje se događaju. Cilj je na temelju tih promjena zaključiti postoji li neki obrazac ponašanja pomoću kojeg možemo opisati buduće vrijednosti. U tu svrhu provodimo analizu vremenskih nizova.

Definicija 1.1 ***Slučajni proces*** *je familija* $\{X_t, t \in T\}, T \subseteq \mathbb{R}$, *slučajnih varijabli definiranih na istom vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, P)$ *.*

Za tako definiran slučajan proces vrijedi sljedeće:

- $\forall t \in T, X_t$ je slučajna varijabla
- $\forall \omega \in \Omega, t \to X_t(\omega)$ je trajektorija slučajnog procesa.

Vremenski niz je niz podataka $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$ prikupljenih u uzastopnim vremenskim trenucima $t_1, t_2, \ldots, t_n$. Svaki podatak vremenskog niza $x_{t_i}, i = 1, \ldots, n$ smatrat ćemo realizacijom jedne slučajne varijable $X_{t_i}, i = 1, \ldots, n$, odnosno vremenski niz $\{X_t, t \in T_0\}$ smatramo (dijelom) realizacije slučajnog procesa $\{X_t, t \in T\}$, u trenucima iz T_0, $T_0 = \{t_1, t_2, \ldots, t_n\}$, $T_0 \subseteq T$. Slučajni proces $\{X_t, t \in T\}$ zovemo model vremenskog niza.
Modelirati vremenski niz znači odrediti u distribucijskom smislu slučajni proces za koji vjerujemo da je taj vremenski niz realizacija ili dio realizacije. S obzirom na to da uvijek raspolažemo samo s dijelom trajektorije slučajnog procesa potrebno je postaviti uvjete na strukturu procesa koje ćemo opisati u nastavku.

1.1 Stacionarni procesi

Definicija 1.2 *Slučajni proces* $\{X_t, t \in T\}$ *je* ***strogo stacionaran*** *(stacionaran u užem smislu) ako* $\forall h > 0$ *vrijedi:*

$$(X_{t_1}, \ldots, X_{t_n}) \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h}), \forall t_1, \ldots, t_n \in T, h \in \mathbb{N}.$$

Posebno, $X_t \stackrel{d}{=} X_s, \forall s, t \in T$.
Nadalje promatramo diskretan T, pri čemu je $T = \mathbb{Z}$.
Strogu stacionarnost je često teško postići te zbog toga provjeravamo druga obilježja procesa kao što su očekivanje, varijanca i kovarijanca. Također važno nam je da se ta obilježja ne mijenjaju tijekom vremena. O tome govori koncept slabe stacionarnosti.

Definicija 1.3 *Za slučajni proces* $\{X_t, t \in \mathbb{Z}\}$ *takav da je* $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$ *definiramo sljedeće funkcije:*

- ***funkcija očekivanja procesa*** X_t *je funkcija* $\mu : \mathbb{Z} \to \mathbb{R}$ *definirana s*

$$\mu(t) = E[X_t]$$

- ***funkcija autokovarijanci (ACVF) procesa*** X_t *je funkcija* $\gamma : \mathbb{Z}^2 \to \mathbb{R}$ *definirana s*

$$\gamma(t,s) = Cov(X_t, X_s) = E[(X_t - E[X_t])(X_s - E[X_s])] = E[X_t X_s] - E[X_t]E[X_s]$$

- ***funkcija autokorelacija (ACRF) procesa*** X_t *je funkcija* $\rho : \mathbb{Z}^2 \to [-1, 1]$ *definirana s*

$$\rho(t,s) = Corr(X_t, X_s) = \frac{Cov(X_t, X_s)}{\sqrt{Var(X_t)Var(X_s)}} = \frac{\gamma(t,s)}{\sqrt{\gamma(t,t)\gamma(s,s)}}$$

Definicija 1.4 *Slučajni proces* $\{X_t, t \in \mathbb{Z}\}$ *je* ***slabo stacionaran*** *(stacionaran u širem smislu) ako vrijedi:*

(i) $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$

(ii) $\mu(t) = c, \forall t \in \mathbb{Z}$ *, gdje je* $c \in \mathbb{R}$ *konstanta*

(iii) $\gamma(t,s) = \gamma(t+h, s+h), \forall t, s, h \in \mathbb{Z}.$

Ako je $\{X_t, t \in \mathbb{Z}\}$ strogo stacionaran proces i ako vrijedi $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$ tada je taj proces i slabo stacionaran.
U nastavku ovog diplomskog rada, stacionaran proces podrazumijevat će proces stacionaran u širem smislu.
Uočimo, ako u (iii) stavimo $s = t$, dobivamo $\gamma(t,t) = \gamma(t+h, t+h)$, odnosno varijanca stacionarnog procesa je konstantna. Nadalje, stavimo li u (iii) $h = -t$, vrijedi $\gamma(t,s) = \gamma(0, s-t), \forall t, s \in \mathbb{Z}$. Dakle, funkcija autokovarijanci ovisi samo o razlici $s - t$, stoga je možemo promatrati kao funkciju jedne varijable. Vrijedi $\gamma(h) = \gamma(0,h) = Cov(X_t, X_{t+h})$.

Propozicija 1.1 (Svojstva funkcije autokovarijanci stacionarnog procesa) *Za funkciju autokovarijanci* γ *stacionarnog procesa* $\{X_t, t \in \mathbb{Z}\}$ *vrijedi:*

(i) $\gamma(0) \geq 0$

(ii) $|\gamma(h)| \leq \gamma(0), \forall h \in \mathbb{Z}$

(iii) $\gamma(h) = \gamma(-h), \forall h \in \mathbb{Z}.$

Dokaz: Prvo svojstvo proizlazi iz činjenice da vrijedi $Var(X_t) \geq 0$, dok je drugo posljedica Cauchy-Schwarz nejednakosti. Posljednje svojstvo se može jednostavno pokazati sljedećim nizom jednakosti:

$$\gamma(h) = Cov(X_{t+h}, X_t) = Cov(X_t, X_{t+h}) = \gamma(-h).$$

□

Analogno navodimo svojstva funkcije autokorelacija stacionarnog procesa koja se također mogu vrlo lako pokazati korištenjem prethodne propozicije.

Propozicija 1.2 (Svojstva funkcije autokorelacija stacionarnog procesa) *Za funkciju autokorelacija ρ stacionarnog procesa $\{X_t, t \in \mathbb{Z}\}$ vrijedi:*

(i) $\rho(0) = 1$

(ii) $|\rho(h)| \leq 1, \forall h \in \mathbb{Z}$

(iii) $\rho(h) = \rho(-h), \forall h \in \mathbb{Z}$.

Neki od najvažnijih primjera stacionarnih procesa su nezavisno jednako distribuirani šum i bijeli šum.

Definicija 1.5 *Neka je $\{X_t, t \in \mathbb{Z}\}$ niz nezavisnih jednako distribuiranih slučajnih varijabli takvih da vrijedi:*

(i) $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$

(ii) $E[X_t] = 0, \forall t \in \mathbb{Z}$

(iii) $Var(X_t) = \sigma^2, \forall t \in \mathbb{Z}$.

Tada je $\{X_t\}$ ***nezavisni jednako distribuirani šum****, oznaka $\{X_t\} \sim IID(0, \sigma^2)$.*

Definicija 1.6 *Neka je $\{X_t, t \in \mathbb{Z}\}$ niz slučajnih varijabli takvih da vrijedi:*

(i) $E[X_t^2] < \infty, \forall t \in \mathbb{Z}$

(ii) $E[X_t] = 0, \forall t \in \mathbb{Z}$

(iii) $Var(X_t) = \sigma^2, \forall t \in \mathbb{Z}$

(iv) $Cov(X_t, X_s) = 0, \forall t \neq s$.

Tada je $\{X_t\}$ ***bijeli šum****, oznaka $\{X_t\} \sim WN(0, \sigma^2)$.*

1.2 Modeli stacionarnih vremenskih nizova

Pretpostavimo da tražimo stacionarni proces kao model za podatke vremenskog niza. Pretpostavit ćemo i da proces ima funkciju očekivanja $\mu(t) = 0, \forall t \in \mathbb{Z}$. Opći okvir za proučavanje stacionarnih procesa nam daju linearni procesi. Svaki stacionaran proces je linearan ili se može transformirati u linearni proces. Taj je rezultat poznat pod nazivom Woldova dekompozicija (vidi [1, Poglavlje 2.6]).

Definicija 1.7 *Kažemo da je* $\{X_t, t \in \mathbb{Z}\}$ ***linearan proces*** *ako se može prikazati kao*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \tag{1.1}$$

gdje je $\{Z_t\} \sim WN(0, \sigma^2)$, $(\psi_j)_{j\in\mathbb{Z}}$ *niz konstanti takav da je* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

Drugim riječima, kažemo da je proces linearan ako se može prikazati kao linearna kombinacija prošlih, sadašnjih i budućih vrijednosti bijelog šuma.

Definicija 1.8 *Za proces* $\{X_t, t \in \mathbb{Z}\}$ ***operator pomaka unazad*** *definiran je s*

$$BX_t = X_{t-1}.$$

Analogno, primjenom j puta, dobivamo $B^j X_t = B^{j-1} X_{t-1} = \cdots = X_{t-j}$.

Linearan proces možemo zapisati kao

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j B^j Z_t = \psi(B) Z_t,$$

gdje je $\psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$ linearan filter.

Propozicija 1.3 *(Linearan filter stacionarnog procesa je stacionaran proces) Neka je* $\{Y_t\}$ *stacionaran proces s očekivanjem 0 i funkcijom autokovarijanci* γ_Y. *Ako je* $\psi(B)$ *linearan filter takav da je* $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$, *tada je proces*

$$X_t = \psi(B) Y_t = \sum_{j=-\infty}^{\infty} \psi_j Y_{t-j}, t \in \mathbb{Z}$$

stacionaran s očekivanjem 0 i funkcijom autokovarijanci

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Y(h + k - j), h \in \mathbb{Z}.$$

Dokaz: Vidi [1, str. 52].

Korolar 1.1 *Linearan proces* $\{X_t, t \in \mathbb{Z}\}$ *je stacionaran s očekivanjem 0 i funkcijom autokovarijanci*

$$\gamma(h) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h}, h \in \mathbb{Z}.$$

Dokaz: Vidi [1, str. 52].

Svaki linearan proces se može prikazati kao

$$X_t = \sum_{j=-\infty}^{0} \psi_j Z_{t-j} + \sum_{j=0}^{\infty} \psi_j Z_{t-j}.$$

Za linearan proces za koji je $\psi_j = 0$ za negativne indekse kažemo da je kauzalan (neovisan o budućnosti). Ako je vremenski niz realizacija stacionarnog procesa, model za podatke može biti $X_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$ za koje je $\psi_j = 0$ za $j > q$, za neki $q \geq 1$. U tu svrhu definiramo sljedeću vrstu procesa.

Definicija 1.9 *Slučajan proces* $\{X_t, t \in \mathbb{Z}\}$ *je* ***proces pomičnih prosjeka reda q*** *ako je*

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, t \in \mathbb{Z},$$

gdje je $\{Z_t\} \sim WN(0, \sigma^2)$, $\theta_1, \ldots, \theta_q \in \mathbb{R}, \theta_q \neq 0, q \in \mathbb{N}$. *Oznaka* $\{X_t\} \sim MA(q)$.

Napomena 1.1

(i) MA(q) je stacionaran proces jer je linearan, $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}, \psi_0 = 1, \psi_j = \theta_j, j = 1, \ldots, q$, $\psi_j = 0$ *za* $j < 0$ *i* $j > q$.

(ii) Prema Korolaru 1.1 funkcija autokovarijanci MA(q) procesa je

$$\gamma(h) = \begin{cases} \sigma^2 \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} & , |h| \leq q \\ 0 & , |h| > q. \end{cases}$$

Kažemo da je stacionaran proces q-**koreliran** ako vrijedi $\gamma(h) = 0$ za $|h| > q, q \in \mathbb{N}$. Naprimjer, bijeli šum je 0-koreliran proces, MA(q) je q-koreliran. Važnost MA(q) procesa je upravo u tome da je svaki q-koreliran proces MA(q). O tome govori sljedeći teorem.

Teorem 1.1 *Ako je* $\{X_t\}$ *stacionaran q-koreliran proces,* $q \in \mathbb{N}$, *s očekivanjem 0, onda se* $\{X_t\}$ *može reprezentirati kao MA(q) proces.*

Dokaz: Vidi [2, str. 89].

Ideja za sljedeću vrstu modela je prikazati sadašnju vrijednost procesa kao linearnu kombinaciju prošlih vrijednosti uz šum.

Definicija 1.10 *Slučajan proces* $\{X_t, t \in \mathbb{Z}\}$ *je* ***autoregresivni proces reda p*** *ako je stacionaran i ako*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, t \in \mathbb{Z}, \tag{1.2}$$

gdje je $\{Z_t\} \sim WN(0, \sigma^2)$, $\phi_1, \ldots, \phi_p \in \mathbb{R}$, $\phi_p \neq 0$. *Oznaka* $\{X_t\} \sim AR(p)$

Izraz (1.2) možemo zapisati kao $\phi(B)Xt = Z_t$, gdje je $\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p, z \in \mathbb{Z}$ karakteristični AR polinom.

Ako se sadašnja vrijednost procesa može zapisati kao linearna kombinacija prošlih vrijednosti procesa te prošlih i sadašnjih vrijednosti šuma, tada govorimo o autoregresivnom pomičnom prosjeku.

Definicija 1.11 *Slučajni proces* $\{X_t, t \in \mathbb{Z}\}$ *je* ***autoregresivni pomični prosjek reda (p,q)****,* $p, q \in \mathbb{R}^+$ *ako je stacionaran i ako je*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, t \in \mathbb{Z}, \tag{1.3}$$

gdje je $\{Z_t\} \sim WN(0, \sigma^2)$, $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q \in \mathbb{R}$, $\phi_p, \theta_q \neq 0$ *i polinomi*

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q$$

nemaju zajedničkih nultočki. Oznaka $X_t \sim ARMA(p, q)$.

Teorem 1.2 *Stacionarno rješenje jednadžbe (1.3) postoji i jedinstveno je ako i samo ako*

$$\phi(z) \neq 0, \ za\ |z| = 1.$$

Odnosno, stacionarno rješenje jednadžbe (1.3) postoji ako i samo ako AR polinom nema jediničnu korijen.

Dokaz: Vidi [2, str. 127].

2 Procjena očekivanja

Stacionaran proces $\{X_t, t \in \mathbb{Z}\}$ karakteriziran je funkcijom očekivanja $\mu(\cdot)$ i funkcijom autokovarijanci $\gamma(\cdot)$. Procjenitelji funkcije očekivanja i autokovarijance, kao i funkcije autokorelacija $\rho(\cdot) = \frac{\gamma(\cdot)}{\gamma(0)}$, imaju važnu ulogu pri zaključivanju o procesu,a posebno u problemu konstruiranja odgovarajućeg modela za podatke. Definirajmo najprije procjenitelja za očekivanje kojeg ćemo u nastavku koristiti za procjenu funkcije autokovarijanci i autokorelacija.
Neka je $\{X_t\}$ stacionaran proces, promatran u trenucima $t = 1, \ldots, n$. Za procjenu očekivanja koristimo **aritmetičku sredinu uzorka** koju definiramo kao

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Aritmetička sredina uzorka je nepristran procjenitelj očekivanja μ stacionarnog procesa $\{X_t\}$. Naime,

$$E[\,\bar{X}_n] \,= \frac{1}{n} \sum_{i=1}^{n} E[\,X_i] \,= \frac{1}{n} n\mu = \mu.$$

Srednje kvadratna greška od $\bar{X}_n$ je:

$$\begin{aligned} E(\bar{X}_n - \mu)^2 &= Var(\bar{X}_n) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i-j=-n}^{n} (n - |i-j|)\gamma(i-j) \\ &= \frac{1}{n} \sum_{h=-n}^{n} \left(1 - \frac{|h|}{n}\right) \gamma(h). \end{aligned}$$

Sljedeći teorem govori o konvergenciji $\bar{X}_n$ u srednje kvadratno smislu prema očekivanju μ, a to povlači i konvergenciju po vjerojatnosti. Zbog toga, osim što je nepristrani procjenitelj za očekivanje, aritmetička sredina uzorka je i konzistentan procjenitelj.

Teorem 2.1 *Neka je $\{X_t\}$ stacionaran proces s očekivanjem μ i funkcijom autokovarijanci γ. Tada za $n \to \infty$ vrijedi:*

(i) Ako $\gamma(h) \to 0$ za $h \to \infty$ onda $Var(\bar{X}_n) = E(\bar{X}_n - \mu)^2 \to 0$

(ii) Ako $\sum_{h=-\infty}^{\infty} |\gamma(h)| < \infty$ onda $nVar(\bar{X}_n) = nE(\bar{X}_n - \mu)^2 \to \sum_{h=-\infty}^{\infty} \gamma(h)$.

Dokaz: Promatramo sljedeće:

$$0 \leq nVar(\bar{X}_n) = \frac{1}{n}\sum_{i,j=1}^{n} Cov(X_i, X_j) = \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h)$$
$$\leq \sum_{|h|<n} |\gamma(h)| = 2\sum_{h=1}^{n} |\gamma(h)| + \gamma(0).$$

Ako $\gamma(h) \to 0$ za $h \to \infty$, tada za svaki $\epsilon > 0$ postoji n_0 takav da $|\gamma(h)| < \frac{\epsilon}{2}$ za $h \geq n_0$. Ako $n > max\{n_0, 2n_0\gamma(0)/\epsilon\}$ tada

$$0 \leq \frac{1}{n}\sum_{h=1}^{n} |\gamma(h)| = \frac{1}{n}\sum_{h=1}^{n_0-1} |\gamma(h)| + \frac{1}{n}\sum_{h=n_0}^{n} |\gamma(h)|$$
$$\leq \frac{n_0\gamma(0)}{n} + \frac{1}{n}(n - n_0)\frac{\epsilon}{2} \leq \frac{n_0\gamma(0)}{n} + \frac{\epsilon}{2} \leq \frac{n_0\gamma(0)\epsilon}{2n_0\gamma(0)} + \frac{\epsilon}{2} = \epsilon.$$

Slijedi da $Var(\bar{X}_n)$ konvergira prema nuli kada $n \to \infty$ što dokazuje prvu tvrdnju. Nadalje, kao dokaz druge tvrdnje promatramo:

$$\lim_{n\to\infty} nVar(\bar{X}_n) = \lim_{n\to\infty} \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h) = \sum_{h=-\infty}^{\infty} \gamma(h) < \infty.$$

$\sum_{h=-\infty}^{\infty} \gamma(h)$ konvergira zbog toga što apsolutno konvergira prema pretpostavci.

□

Za donošenje zaključaka o očekivanju, koristeći uzoračku aritmetičku sredinu, potrebno je poznavati ili aproksimirati distribuciju od $\bar{X}_n$. Ako je slučajni proces Gaussovski tada vrijedi

$$n^{\frac{1}{2}}(\bar{X}_n - \mu) \sim \mathcal{N}\left(0, \sum_{|h|<n}\left(1 - \frac{|h|}{n}\right)\gamma(h)\right).$$

Za mnoge vremenske nizove, a posebno za linearne i ARMA modele, distribucija od $\bar{X}_n$ je približno jednaka normalnoj distribuciji s očekivanjem μ i varijancom $n^{-1}\sum_{|h|<\infty}\gamma(h)$, za veliki n. Tada je 95% pouzdani interval za očekivanje

$$\left(\bar{X}_n - \frac{1.96v^{\frac{1}{2}}}{\sqrt{n}}, \bar{X}_n + \frac{1.96v^{\frac{1}{2}}}{\sqrt{n}}\right),$$

pri čemu je $v = \sum_{|h|<\infty}\gamma(h)$. Općenito v nije poznat i trebamo ga procijeniti. Kako bi procijenili v potrebno je procijeniti funkciju autokovarijanci, o tome govorimo u sljedećem poglavlju.

3 Uzoračka funkcija autokorelacija

Definicija 3.1 *Za niz slučajnih varijabli $X_1, X_2, \ldots, X_n$* ***uzoračka funkcija autokovarijanci*** *je definirana s*

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{j=1}^{n-h}(X_{j+h} - \bar{X}_n)(X_j - \bar{X}_n), \quad za \quad 0 \leq h < n,$$

i $\widehat{\gamma}(h) = \widehat{\gamma}(-h)$, za $-n < h < 0$.
Uzoračka funkcija autokorelacija *je zadana s*

$$\widehat{\rho}(h) = \frac{\widehat{\gamma}(h)}{\widehat{\gamma}(0)}, \quad za \quad -n < h < n.$$

Napomena 3.1

(i) Uočimo da je $\widehat{\gamma}(h)$ za $h > 0$ približno jednaka uzoračkoj kovarijanci $n-h$ parova $(X_1, X_{1+h}), (X_2, X_{2+h}), \ldots$ razlika se pojavljuje pri dijeljenju s n (a ne s $n-h$) i oduzimanju prosjeka svih.

(ii) Procjena ima smisla samo za stacionarne procese jer su inače γ i ρ funkcije dvije varijable.

(iii) Za velike h (blizu n) postoji malo podataka, (n-h) parova, iz tog razloga preporuka je koristiti procjenu u slučaju kad imamo barem 50 podataka i kada je $h \leq \frac{n}{4}$ (prema [1])

Procjenitelji $\widehat{\gamma}(h)$ i $\widehat{\rho}(h)$ nisu nepristrani za $\gamma(h)$ i $\rho(h)$ zbog toga što je suma normalizirana (dijeljena) sa n, a ne sa $n-h$. Međutim, za velike uzorke $\widehat{\gamma}(h)$ i $\widehat{\rho}(h)$ su približno nepristrani.

Propozicija 3.1 *Uzoračka funkcija autokovarijanci ima svojstvo da za svaki $k \geq 1$ k-dimenzionalna uzoračka matrica kovarijanci*

$$\widehat{\mathbf{\Gamma}}_{\mathbf{k}} = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \ldots & \widehat{\gamma}(k-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \ldots & \widehat{\gamma}(k-2) \\ \vdots & \vdots & \ldots & \vdots \\ \widehat{\gamma}(k-1) & \widehat{\gamma}(k-2) & \ldots & \widehat{\gamma}(0) \end{pmatrix}$$

je pozitivno semidefinitna.

Dokaz: Ako je $\widehat{\Gamma}_n$ pozitivno semidefinitna, tada je $\widehat{\Gamma}_k$ pozitivno semidefinitna za sve $k < n$. Pretpostavimo da je $k \geq n$ i zapišimo

$$\widehat{\Gamma}_k = n^{-1}TT',$$

gdje je T $k \times 2k$ matrica

$$T = \begin{pmatrix} 0 & \dots & 0 & 0 & Y_1 & Y_2 & \dots & Y_k \\ 0 & \dots & 0 & Y_1 & Y_2 & \dots & Y_k & 0 \\ \vdots & & & & & & & \vdots \\ 0 & Y_1 & Y_2 & \dots & Y_k & 0 & \dots & 0 \end{pmatrix},$$

pri čemu je $Y_i = X_i - \bar{X}_n$ za $i = 1, \dots, n$ i $Y_i = 0$ za $i = n+1, \dots, k$. Tada za svaki $k \times 1$ vektor $\mathbf{a}$ vrijedi

$$\mathbf{a}'\widehat{\Gamma}_k\mathbf{a} = n^{-1}(\mathbf{a}'T)(T'\mathbf{a}) \geq 0.$$

Slijedi da su uzoračka matrica kovarijanci $\widehat{\Gamma}_k$ i uzoračka matrica korelacija

$$\widehat{R}_k = \frac{\widehat{\Gamma}_k}{\widehat{\gamma}(0)}$$

pozitivno semidefinitne. □

Propozicija 3.2 *Matrice $\widehat{\Gamma}_k$ i $\widehat{R}_k$ su regularne ako postoji barem jedan Y_i različit od* 0 *ili ako je $\widehat{\gamma}(0) > 0$.*

Dokaz: Pretpostavimo da je $\widehat{\gamma}(0) > 0$ i $\widehat{\Gamma}_k$ singularna. Tada za vektor $\mathbf{a}$ različit od nula slijedi da je $\mathbf{a}'T = 0$ i rang od T je manji od k. Neka je Y_i prvi u nizu $Y_1, Y_2, \dots, Y_k$ različit od nule i promotrimo $k \times k$ submatricu od T koja sadrži stupce od $i+1$ do $i+k$. Ta matrica je donjetrokutasta s elementima na dijagonali jednakima Y_i te je stoga apsolutna vrijednost determinante jednaka $|Y_i|^k \neq 0$. Slijedi da je submatrica nesingularna i da je T ranga k, pri čemu dolazimo do kontradikcije. □

Distribuciju od $\widehat{\rho}(h)$ tipično ne znamo, ali ju možemo dobro aproksimirati normalnom distribucijom za velike uzorke. Za linearne modele, specijalno za ARMA modele, $\widehat{\rho}_k = (\widehat{\rho}(1), \dots, \widehat{\rho}(k))'$ je asimptotski distribuirana, za velike n, kao $\mathcal{N}(\rho_k, n^{-1}W)$. To jest

$$\widehat{\rho} \overset{a}{\sim} \mathcal{N}(\rho, n^{-1}W),$$

pri čemu je $\rho = (\rho(1), \dots, \rho(k))'$ i W matrica kovarijanci s (i,j) elementom koji je zadan Bartlettovom formulom:

$$w_{ij} = \sum_{k=-\infty}^{\infty} \{\rho(k+i)\rho(k+j) + \rho(k-i)\rho(k+j) + 2\rho(i)\rho(j)\rho^2(k) - 2\rho(i)\rho(k)\rho(k+j) - 2\rho(j)\rho(k)\rho(k+i)\}.$$

w_{ij} možemo zapisati i u sljedećem obliku:

$$w_{ij} = \sum_{k=1}^{\infty} \{\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)\} \times \{\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)\}.$$

U nastavku pogledajmo svojstvo sume uzoračke funkcijske autokorelacija koje se koristi u spektralnoj analizi.

Teorem 3.1 *Suma uzoračke funkcije autokorelacija, S_{acf}, na koraku $h \geq 1$ za svaki stacionarni proces je uvijek $\frac{-1}{2}$, to jest*

$$S_{acf} = \sum_{h=1}^{n-1} \widehat{\rho}(h) = -\frac{1}{2}, \quad n \geq 2$$

Dokaz:

$$\begin{aligned} S_{acf} &= \sum_{h=1}^{n-1} \widehat{\rho}(h) = \sum_{h=1}^{n-1} \left(\frac{\sum_{h=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{h=1}^{n} (y_t - \bar{y})^2} \right) \\ &= \frac{\sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2} \\ &= \frac{\sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})}{(\sum_{t=1}^{n} (y_t - \bar{y}))^2 - 2\sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})} \\ &= \frac{\sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})}{-2\sum_{h=1}^{n-1} \sum_{t=1}^{n-h} (y_t - \bar{y})(y_{t+h} - \bar{y})} \\ &= -\frac{1}{2}. \end{aligned}$$

□

Napomena 3.2

(i) S_{acf} ne ovisi o duljini n; $S_{acf} = -\frac{1}{2}, \quad za \quad n \geq 2.$

(ii) Vrijednosti od $\widehat{\rho}(h)$ su linearno zavisne i vrijedi:

$$\widehat{\rho}(i) = -\frac{1}{2} - \sum_{j \neq i} \widehat{\rho}(j), \quad i = 1, \ldots, n-1.$$

(iii) Postoji barem jedna negativna vrijednost od $\widehat{\rho}(h)$ za svaki stacionaran proces.

3.1 Primjeri

Primjer 3.1 Neka je $X_t \sim IID(0, \sigma^2)$, tada je $\rho(h) = 0$ za $|h| > 0$. Prisjetimo se da za linearne modele vrijedi da je $\widehat{\rho} \overset{a}{\sim} \mathcal{N}(\rho, n^{-1}W)$, pri čemu je $\widehat{\rho} = (\widehat{\rho}(1), \ldots, \widehat{\rho}(k))'$. Elemente matrice kovarijanci W računamo prema Bartlettovoj formuli i za ovaj primjer dobivamo da vrijedi

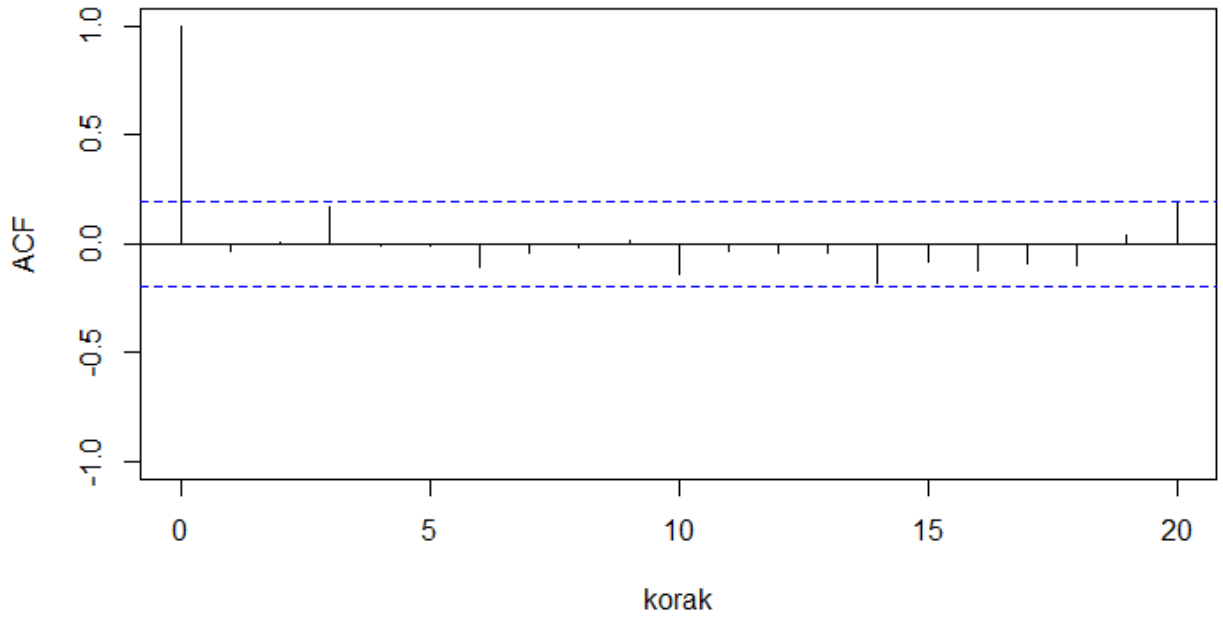
$$w_{ij} = \begin{cases} 1 & , za \quad i = j \\ 0 & , inače. \end{cases}$$

Stoga, za veliki n, $\widehat{\rho}(1), \ldots, \widehat{\rho}(h)$ su asimptotski nezavisne i jednako distribuirane normalnom distribucijom s očekivanjem 0 i varijancom n^{-1}. Na temelju toga možemo konstruirati 95% pouzdani interval, dobivamo da su granice tog pouzdanog intervala $\pm 1.96 n^{-\frac{1}{2}}$.
Slika 3.1 prikazuje uzoračku funkciju autokorelacija za $WN(0, 1)$ veličine uzorka 100. Granice 95% intervala su ± 0.196. Primjećujemo da se svi procijenjeni autokorelacijski koeficijenti nalaze unutar granica pouzdanog intervala što je jedna od glavnih karakteristika za zaključivanje da promatrani proces možemo reprezentirati bijelim šumom.
Primijetimo sljedeće, ako promatramo slučajni proces za koji ne znamo funkciju autokorelacija, umjesto da ispitujemo svaki koeficijent korelacije posebno, možemo testirati ujedinjenu hipotezu da su svi korelacijski koeficijenti do reda k jednaki 0, odnosno da vrijedi $\rho(1) = \rho(2) = \cdots = \rho(k) = 0, k = 1, 2, \ldots$. Kako $\sqrt{n}\widehat{\rho}(h)$ ima asimptotski standardnu normalnu distribuciju, tada suma kvadriranih procijenjenih autokorelacijskih koeficijenata ima asimptotsku χ^2 distribuciju s k stupnjeva slobode. Ova test-statistika naziva se Box-Pierce statistika:

$$Q = n \sum_{h=1}^{k} \widehat{\rho}^2(h) \sim \chi^2(k).$$

Možemo koristiti i Ljung-Box statistiku:

$$Q' = n(n+2) \sum_{h=1}^{k} \frac{\widehat{\rho}^2(h)}{n-h} \sim \chi^2(k).$$



Slika 3.1: Uzoračka funkcija autokorelacija za WN(0,1)

Primjer 3.2 Neka je $X_t \sim MA(1)$ proces

$$X_t = Z_t + \theta Z_{t-1}, t \in \mathbb{Z},$$

gdje je $Z_t \sim IID\,(0, \sigma^2)$. Prema *Korolaru 1.1* vrijedi:

$$\gamma(h) = \begin{cases} \sigma^2(1+\theta^2) & \text{, za} \quad h = 0 \\ \sigma^2\theta & \text{, za} \quad h = \pm 1 \\ 0 & \text{, za} \quad |h| > 1. \end{cases}$$

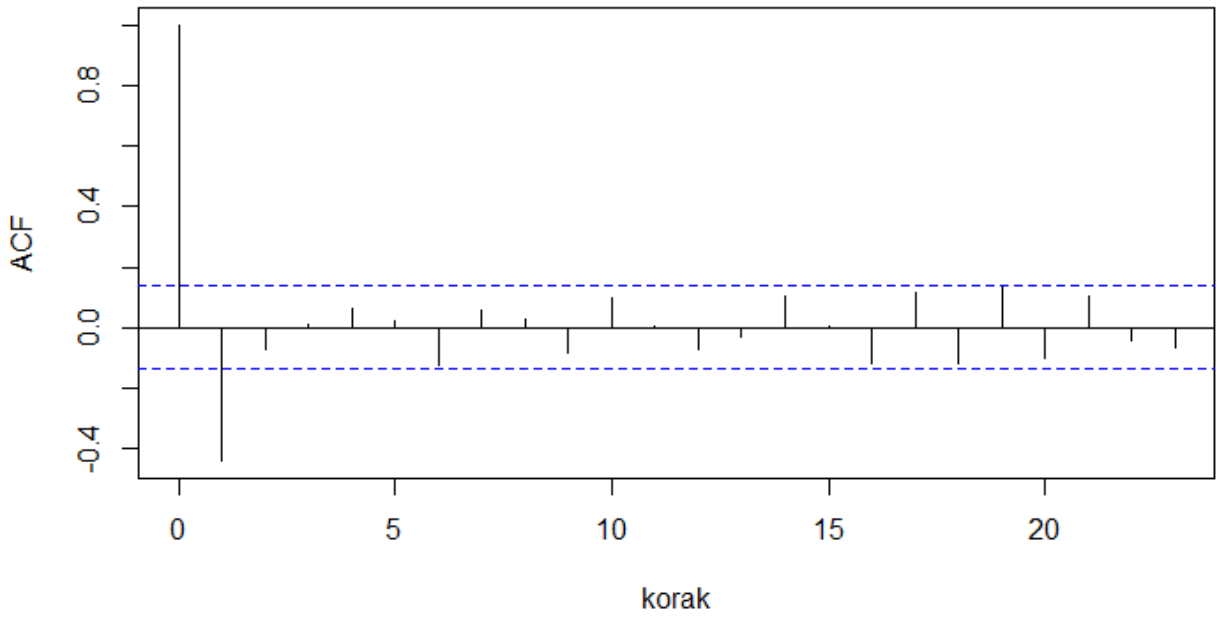
Slijedi da je funkcija autokorelacija zadana s:

$$\rho(h) = \begin{cases} 1 & \text{, za} \quad h = 0 \\ \theta/(1+\theta^2) & \text{, za} \quad h = \pm 1 \\ 0 & \text{, za} \quad |h| > 1. \end{cases}$$

Matrica kovarijanci W je određena s

$$w_{ii} = \begin{cases} 1 - 3\rho^2(1) + 4\rho^4(1) & \text{, za} \quad i = 1 \\ 1 + 2\rho^2(1) & \text{, za} \quad i > 1. \end{cases}$$

Slika 3.2 prikazuje uzoračku funkciju autokorelacija za proces $X_t = Z_t - 0.8Z_{t-1}$ veličine uzorka $n = 200$. Granice 95%-tnog pouzdanog intervala dane s $\pm 1.96 n^{-\frac{1}{2}}[1 + 2\rho(1)^2]^{\frac{1}{2}} = \pm 0.1684$. Očito je da samo na koraku 1 uzoračka funkcija autokorelacija ne upada u pouzdani interval. To sugerira da se radi o MA(1) procesu.



Slika 3.2: Uzoračka funkcija autokorelacija za MA(1) proces $X_t = Z_t - 0.8Z_{t-1}, t \in \mathbb{Z}$

Primjer 3.3 Neka je $X_t \sim AR(1)$ proces,

$$X_t = \phi X_{t-1} + Z_t, t \in \mathbb{Z} \tag{3.2}$$

gdje je $Z_t \sim IID(0, \sigma^2)$ i $|\phi| < 1$. Primijetimo da vrijedi $E(X_t) = 0$. Kako bi dobili funkciju autokovarijanci, jednadžbu (3.2) pomnožimo s obje strane s X_{t-h} i djelujemo s očekivanjem. Slijedi

$$\begin{aligned}\gamma(h) &= Cov(X_t, X_{t-h}) \\ &= Cov(\phi X_{t-1}, X_{t-h}) + Cov(Z_t, X_{t-h}) \\ &= \phi\gamma(h-1) + 0 = \cdots = \phi^h \gamma(0).\end{aligned}$$

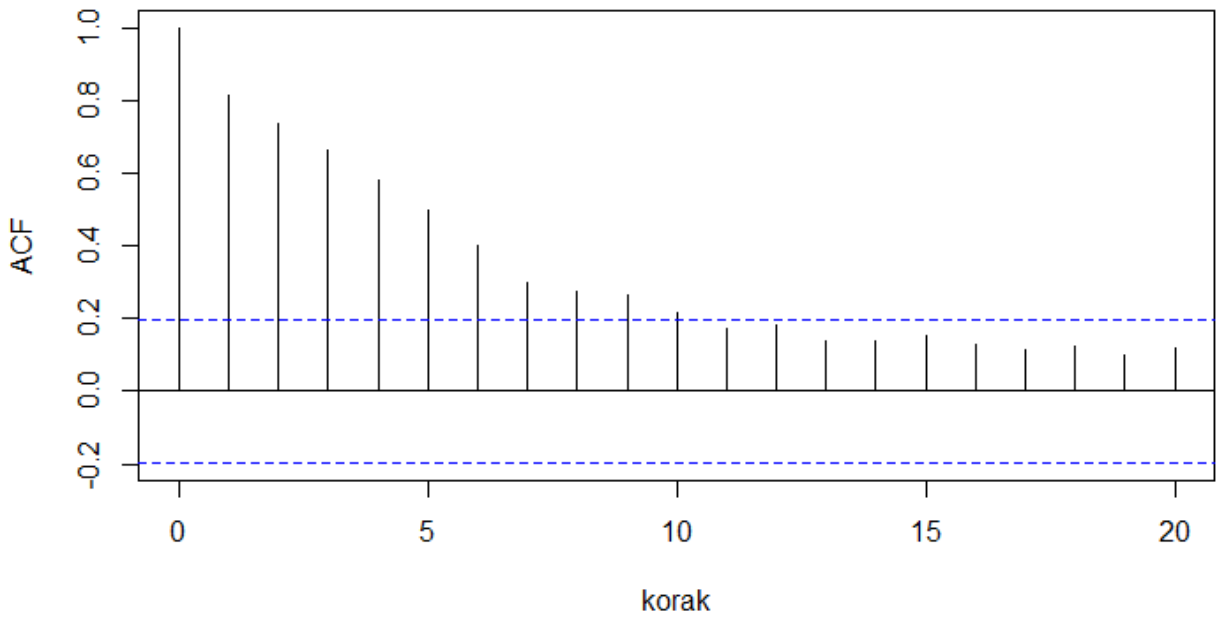
Nadalje,

$$\begin{aligned}\gamma(0) &= Cov(X_t, X_t) \\ &= Cov(\phi X_{t-1} + Z_t, \phi X_{t-1} + Z_t) \\ &= \phi^2\gamma(0) + \sigma^2\end{aligned}$$

Dobivamo da vrijedi $\gamma(0) = \sigma^2/(1-\phi^2)$. Funkcija autokorelacija je stoga jednaka $\rho(h) = \phi^{|h|}$. Matrica kovarijanci W određena je s

$$\begin{aligned}w_{ii} &= \sum_{k=1}^{i} \phi^{2i}(\phi^{-k} - \phi^k)^2 + \sum_{k=i+1}^{\infty} \phi^{2k}(\phi^{-i} - \phi^i)^2 \\ &= \frac{(1-\phi^{2i})(1+\phi^2)}{1-\phi^2} - 2i\phi^{2i}.\end{aligned}$$

Slika 3.3 prikazuje uzoračku funkciju autokorelacija za proces za $X_t = 0.8X_{t-1} + Z_t$ veličine uzorka 100. Koeficijenti autokorelacije eksponencijalno opadaju što je karakteristika AR(1) procesa. Za AR modele puno je prikladnije promatrati parcijalnu funkciju autokorelacija.



Slika 3.3: Uzoračka funkcija autokorelacija za AR(1) proces $X_t = 0.8X_{t-1} + Z_t, t \in \mathbb{Z}$

4 Robusni procjenitelji autokorelacijske funkcije

U prethodnom poglavlju naveli smo klasičnog procjenitelja autokorelacijske funkcije. Međutim, taj procjenitelj nije robustan što znači da samo jedna stršeća vrijednost (engl. *outlier*) može narušiti i sakriti pravu strukturu koreliranosti. Nije rijetkost da više od 10% podataka čine stršeće vrijednosti i kako one često ostaju nezamijećene to predstavlja ozbiljan problem. Zbog toga navodimo robusne autokorelacijske procjenitelje temeljene na konstruiranju dijagrama raspršenja i primjeni robusne regresije. Promatrat ćemo tri najčešće robusne regresijske tehnike.

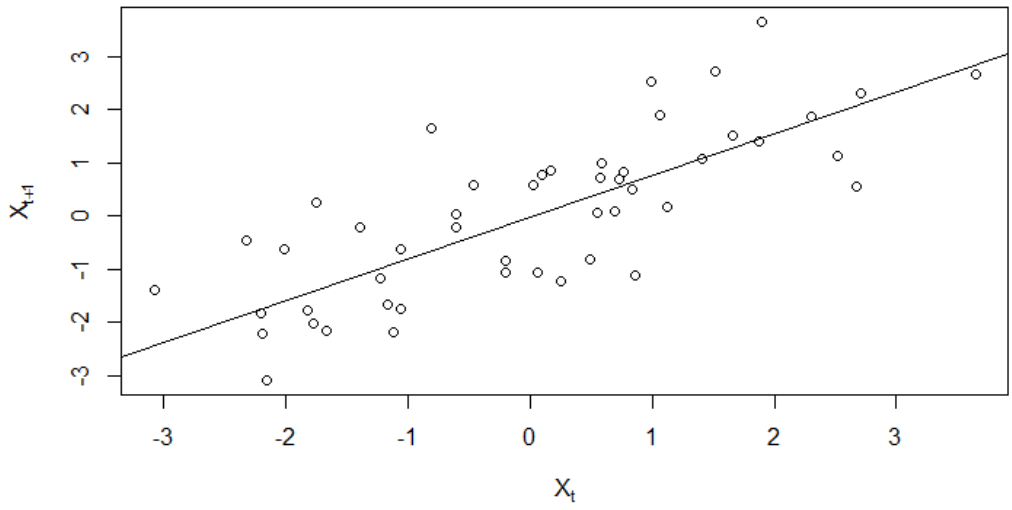
Neka je $\{X_t\}$ stacionaran proces i $X_1, \ldots, X_n$ niz realizacija tog procesa. Fiksiramo korak $h < n, h \in \mathbb{Z}^+$ i promatramo dijagram raspršenja parova $\{(X_t - \bar{X}_n, X_{t+h} - \bar{X}_n), t \in \{1, \ldots, n-h\}\}$.
Ako je $\{X_t\}$ AR(p) proces, $p \in \mathbb{N}$,

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t,$$

gdje je $Z_t \sim IID(0, \sigma^2)$, tada je

$$E[X_{t+h} - \mu | X_t] = (X_t - \mu)\rho(h).$$

Ako se ne radi o regularnom AR(p) ili AR(∞) procesu s IID distribucijom šuma, tada prethodna jednakost vrijedi ako spektralna gustoća od X_t postoji i pozitivna je. U tom slučaju $E[Y|X]$ označava ortogonalnu projekciju Y na linearan prostor razapet s X. Naime, uvjetno očekivanje od X_{t+k} temeljeno na prošlosti $X_t, X_{t-1}, \ldots$ ($E[X_{t+k}|X_s, s \leq t]$) je najbolji linearni prediktor $\mathcal{P}_{\mathcal{M}_t}(X_{t+k})$, pri čemu je k pozitivan, $\mathcal{P}_{\mathcal{C}}$ označava ortogonalnu projekciju na $\mathcal{C}$, a $\mathcal{M}_{\mathcal{C}} = \overline{span}\{X_j, j \leq s\}$ najmanji zatvoren potprostor generiran slučajnim varijablama $\{X_j, j \leq s\}$ (vidi [6]) .



Slika 4.1: Dijagram raspršenja (X_t, X_{t+1}) za realizaciju duljine 50 AR(1) procesa,$\phi = 0.8$

Ako promatramo točku $(X_t - \bar{X}_n, X_{t+h} - \bar{X}_n)$ na dijagramu raspršenja, ona ima tendenciju da se nalazi blizu pravca $y = \rho(h)x$. Za očekivati je da je nagib regresijskog pravca razuman procjenitelj funkcije autokorelacije $\rho(h)$. To sugerira i Slika 4.1.
Prisjetimo se metode najmanjih kvadrata (engl. *Ordinary Least Squares*) koja se temelji na minimiziranju sume kvadrata odstupanja teorijskih od eksperimentalnih vrijednosti. Pokažimo da je procjenitelj dobiven OLS regresijskom tehnikom skoro identičan uzoračkoj funkciji autokorelacija za mali $\frac{h}{n}$. Ako točke dijagrama raspršenja označimo s (x_i, y_i), pri čemu je $y_i = X_{i+h}$, $x_i = X_i$ tada imamo

$$\begin{aligned}
\widehat{\rho}_{OLS}(h) &= \frac{\sum_{j=1}^{n-h}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n-h}(x_j - \bar{x})^2} \\
&= \frac{\sum_{j=1}^{n-h}(X_{j+h} - \bar{X}_{(h+1)\dots n})(X_j - \bar{X}_{1\dots(n-h)})}{\sum_{j=1}^{n-h}(X_j - \bar{X}_{1\dots(n-h)})^2} \\
&\approx \frac{\sum_{j=1}^{n-h}(X_{j+h} - \bar{X})(X_j - \bar{X})}{\frac{n-h}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2} \\
&= \frac{n}{n-h}\widehat{\rho}(h),
\end{aligned}$$

pri čemu je $\bar{x}_{a\dots b} := (b - a + 1)^{-1}\sum_{j=a}^{b} x_j$ i $\bar{x} := \bar{x}_{1\dots n}$.
Zbog faktora $\frac{n}{n-h}$ nagib regresijskog pravca je nepristran procjenitelj za razliku od uzoračke funkcije autokorelacije $\widehat{\rho}(h)$.

Ako provodimo robusnu linearnu regresiju za parove $\{X_t, X_{t+h}\}$ dobivamo robusnog procjenitelja autokorelacije. Definiramo $\widehat{\rho}_{robust}(h)$ kao procjenitelja od β_1 u linearnoj regresiji

$$X_t = \beta_0 + \beta_1 X_{t-h} + \epsilon_t, \tag{4.1}$$

gdje ϵ_t označava grešku.
U nastavku ćemo detaljnije opisati tri različita procjenitelja, temeljena na tri robusne regresijske metode: L1 metoda, metoda najmanjih odrezanih kvadrata i MM metoda.

Procjenitelje definiramo u okviru klasičnog linearnog modela, pretpostavljamo da među podacima postoji linearna veza u obliku sljedećih jednadžbi

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + e_i, \quad i = 1, \ldots, n,$$

gdje je n veličina uzorka. Varijable $X_1, \ldots, X_p$ kojima modeliramo podatke $x_{i1}, \ldots, x_{ip}$ nazivamo nezavisne varijable (prediktor ili egzogena varijabla), a varijablu Y kojom modeliramo y_i nazivamo zavisna varijabla (endogena varijabla). U klasičnoj teoriji za greške modela ϵ_i se pretpostavlja da su međusobno nezavisne te normalno distribuirane s očekivanjem 0 i nepoznatom standardnom devijacijom σ. Podatke možemo zapisati kao vektor redak dimenzije $p+1$ $(\boldsymbol{x}_i, y_i) = (x_{i1}, \ldots, x_{ip}, y_i)$, a nepoznati parametar $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ kao p-dimenzionalni vektor stupac. Primjenom regresijske procjene na podatke dobivamo

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_p, \end{bmatrix}$$

pri čemu su $\widehat{\theta}_j$, $j = 1, \ldots, p$ regresijski koeficijenti. Stvarni θ_j su nepoznati, no uvrštavanjem regresijskih koeficijenata u jednadžbu dobivamo

$$\widehat{y}_i = \widehat{\theta}_1 x_{i1} + \cdots + \widehat{\theta}_p x_{ip} + e_i, \quad i = 1, \ldots, n,$$

$\widehat{y}i$ nazivamo teorijske vrijednosti ili predikcije od y_i. Rezidual r_i i-tog podatka je razlika između stvarne i procijenjene vrijednosti, to jest $r_i = y_i - \widehat{y}_i$, $i = 1, \ldots, n$.

U našem slučaju kada govorimo o procjenitelju autokorelacijske funkcije definiranog s (4.1) radi se o jednostavnoj linearnoj regresiji, to jest $p = 2$, $x_{i1} := x_i$ i $x_{i2} := 1$, za $i = 1, \ldots, n$.

Kako bi mogli napraviti usporedbu među procjeniteljima potrebno je uvesti određenu mjeru robusnosti. U tu svrhu definiramo prijelomnu vrijednost (engl. *breakdown point*). Pretpostavimo da imamo uzorak od n točaka podataka

$$Z = \{(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, \ldots, x_{np}, y_n)\},$$

neka je T regresijski procjenitelj. Djelovanjem s T na uzorak Z dobivamo vektor regresijskih koeficijenata

$$T(Z) = \widehat{\boldsymbol{\theta}}.$$

Promotrimo sve moguće uzorke Z' koje dobijemo tako da zamijenimo bilo kojih m točaka iz originalnog uzorka s proizvoljnim vrijednostima. Definiramo s $b(m; T, Z)$ maksimalno odstupanje koje može biti uzrokovano takvom zamjenom, to jest

$$b(m; T, Z) = \sup_{Z'} \|T(Z') - T(Z)\|.$$

Ako je $b(m; T, Z)$ beskonačno, to znači da m stršećih vrijednosti može imati proizvoljno velik utjecaj na T. Stoga, **prijelomnu vrijednost** procjenitelja T na uzorku Z možemo definirati kao

$$\epsilon_n^*(T, Z) = min\Big\{\frac{m}{n} : b(m; T, Z) = \infty\Big\}.$$

Drugim riječima, to je najmanji omjer kontaminiranosti podataka koji će uzrokovati da procjenitelj T poprimi vrijednosti proizvoljno daleko od $T(Z)$. Primijetimo da ova definicija ne ovisi o vjerojatnosnim distribucijama. Često se promatra i asimptotska prijelomna vrijednost koja je limes prijelomne vrijednosti na konačnom uzorku kad veličina uzorka teži u beskonačno.
Na primjer, za metodu najmanjih kvadrata dovoljna je jedna stršeća vrijednost da bi vrijednost procjene T bila proizvoljno daleko. Prijelomna vrijednost je

$$\epsilon_n^*(T, Z) = \frac{1}{n},$$

što konvergira u 0 kad veličina uzorka teži u beskonačno te je zbog toga asimptotska prijelomna vrijednost procjenitelja $\widehat{\rho}_{OLS}$ 0%.

Osim prijelomne vrijednost, kao mjeru robusnosti promatrat ćemo i asimptotsku efikasnost. Neka je $\widehat{\rho}$ procjenitelj koji je asimptotski normalan s matricom kovarijanci $\boldsymbol{V}$. Neka je $\tilde{\rho}$ MLE procjenitelj (procjenitelj metodom maksimalne vjerodostojnosti) s asimptotskom matricom kovarijanci $\boldsymbol{V}_0$. Za vektor $\boldsymbol{c} \in \mathbb{R}^p$ asimptotska varijanca linearnih kombinacija $\boldsymbol{c}'\widehat{\rho}$ i $\boldsymbol{c}'\tilde{\rho}$ je redom $\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}$ i $\boldsymbol{c}'\boldsymbol{V}_0\boldsymbol{c}$. **Asimptotska efikasnost** procjenitelja $\widehat{\rho}$ je definirana kao

$$eff(\widehat{\rho}) = \min_{\boldsymbol{c} \neq \boldsymbol{0}} \frac{\boldsymbol{c}'\boldsymbol{V}_0\boldsymbol{c}}{\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c}}.$$

4.1 L1 procjenitelj

Kako smo već spomenuli, najčešće korišteni regresijski procjenitelj kojeg su predstavili Gauss[1] i Legendre[2] je procjenitelj temeljen na metodi najmanjih kvadrata

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} r_i^2.$$

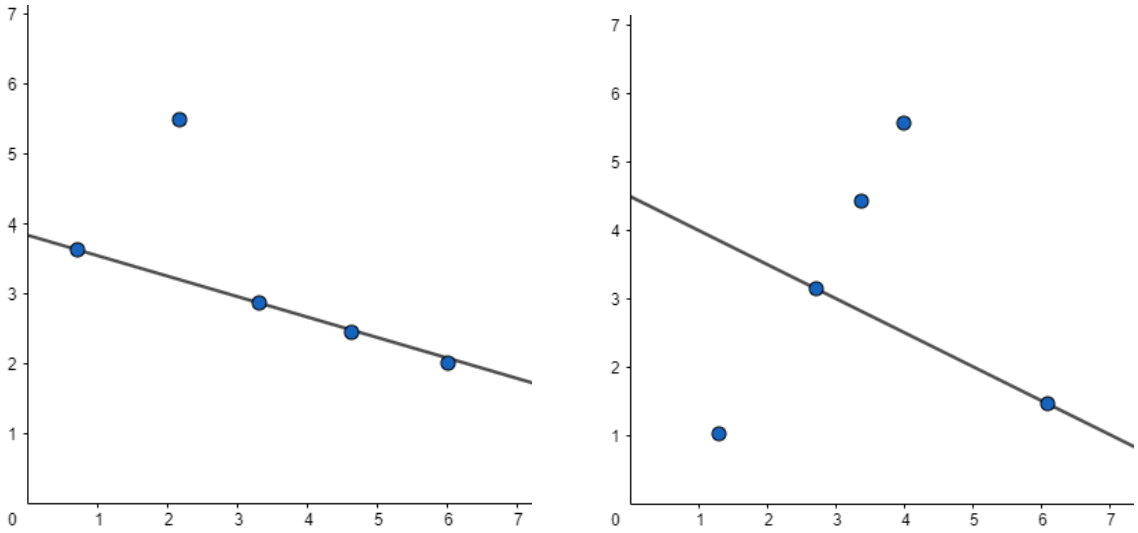
Prvi korak prema izgradnji robusnog procjenitelja kao poboljšanje OLS regresijske metode napravio je Edgeworth[3]. Tvrdio je da stršeće vrijednosti imaju velik utjecaj na OLS metodu zato što su reziduali kvadrirani. Zbog toga predlaže LAD (engl. *least absolute deviations*) ili **L1 procjenitelja** koji se dobije kao

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} |r_i|.$$

L1 regresija u jednodimenzionalnom slučaju ($p = 1$) kao rješenje daje median uzorka.
Iako je asimptotska prijelomna vrijednost univarijatnog mediana 50%, asimptotska prijelomna vrijednost L1 regresije je 0%. Objasnimo to na sljedećem primjeru.

Primjer 4.1 Pretpostavimo da je zadano pet točaka u ravnini $(x_1, y_1), \ldots, (x_5, y_5)$ i da je iz nekog razloga vrijednost y_2 izmjerena kao puno veća. Takvu točku nazivamo stršeća vrijednost u y smjeru i ona nema utjecaj na L1 pravac, što prikazuje Slika 4.2 (a). To predstavlja značajan pomak u odnosu na OLS regresiju i kažemo da je L1 pravac otporan na stršeće vrijednosti u y smjeru. Nadalje, ako je iz nekog razloga točka x_2 izmjerena kao puno veća, tada ona predstavlja stršeću vrijednost u x smjeru. Takva točka ima velik utjecaj na L1 regresijski pravac kao što je prikazano na Slici 4.2 (b). Stoga, L1 regresija je u općem slučaju osjetljiva na stršeće vrijednosti i njena asimptotska prijelomna vrijednost je 0%.

[1] Carl Friedrich Gauss (rođen 1777.g.),njemački matematičar i astronom
[2] Adrien-Marie Legendre (rođen 1752.g.), francuski matematičar i astronom
[3] Francis Ysidro Edgeworth (rođen 1845.g.),anglo-irski filozof i politički ekonomist



(a) stršeća vrijednost u y smjeru (b) stršeća vrijednost u x smjeru

Slika 4.2: L1 regresijski pravac

4.2 Procjenitelj metodom najmanjih odrezanih kvadrata

Zbog loše prijelomne vrijednost L1 procjenitelja, kao daljnji napredak regresijskih metoda razmatralo se o ideji da se suma zamijeni medijanom koji je robustan procjenitelj. Predloženi su LMS procjenitelji (engl. *least median of squares*)

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \operatorname*{med}_i r_i^2.$$

Pokazalo se da LMS metoda ima vrlo lošu asimptotsku efikasnost te kako bi se to popravilo Rousseeuw [4] je uveo procjenitelja **metodom najmanjih odrezanih kvadrata (LTS procjenitelj)** (engl. *least trimmed squares*) dobivenog kao

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{h} r_{i:n}^2, \quad \text{za neki } h \leq n,$$

gdje su $(r^2)_{1:n} \leq \cdots \leq (r^2)_{n:n}$ poredani kvadrirani reziduali (primijetimo da su reziduali prvo kvadrirani pa onda poredani po veličini). O odabiru h govorit ćemo u nastavku rada. LTS metoda je vrlo slična OLS metodi, iz sume su samo izbačeni reziduali s najvećom kvadriranom vrijednosti. U našem slučaju minimizirat ćemo sumu najmanjih 50% reziduala.

Važno svojstvo procjenitelja, koje je za razliku od L1 procjenitelja zadovoljeno za LTS procjenitelja, je ekvivarijantnost. Možemo promatrati tri tipa ekvivarijantnosti, koji su redom

[4] Peter Rousseeuw (rođen 1956.g.), belgijski statističar

po važnosti: regresijska ekvivarijantnost, ekvivarijantnost na skaliranje i afina ekvivarijantnost.
Kažemo da je procjenitelj T **regresijski ekvivarijantan** ako

$$T(\{(\boldsymbol{x}_i, y_i + \boldsymbol{x}_i\boldsymbol{v}); i = 1, \ldots, n\}) = T(\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}) + \boldsymbol{v},$$

gdje je $\boldsymbol{v}$ bilo koji vektor stupac.
Kažemo da je procjenitelj T **ekvivarijantan na skaliranje** ako

$$T(\{(\boldsymbol{x}_i, cy_i); i = 1, \ldots, n\}) = cT(\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}),$$

gdje je c neka konstanta. Ovo svojstvo govori da je regresijska procjena neovisna o izboru mjerne jedinice varijable y.
Kažemo da je procjenitelj T **afino ekvivarijantan** ako

$$T(\{(\boldsymbol{x}_i A, y_i); i = 1, \ldots, n\}) = A^{-1}T(\{(\boldsymbol{x}_i, y_i); i = 1, \ldots, n\}),$$

za svaku regularnu matricu A. Govori da linearna transformacija vektora $\boldsymbol{x}_i$ uzrokuje i odgovarajuću promjenu procjenitelja T.

Lema 4.1 *LTS procjenitelj je regresijski ekvivarijantan, ekvivarijantan na skaliranje i afino ekvivarijantan.*

Dokaz: Tvrdnja slijedi iz sljedećih jednakosti:

$$\sum_{i=1}^{h}((y_i + \boldsymbol{x}_i\boldsymbol{v} - \boldsymbol{x}_i(\boldsymbol{v} + \boldsymbol{\theta}))^2)_{i:n} = \sum_{i=1}^{h}((y_i - \boldsymbol{x}_i\boldsymbol{\theta})^2)_{i:n} = \sum_{i=1}^{h} r^2_{i:n},$$

$$\sum_{i=1}^{h}((cy_i - \boldsymbol{x}_i(c\boldsymbol{\theta}))^2)_{i:n} = c^2\sum_{i=1}^{h}((y_i - \boldsymbol{x}_i\boldsymbol{\theta})^2)_{i:n},$$

$$\sum_{i=1}^{h}((y_i - \boldsymbol{x}_i A(A^{-1}\boldsymbol{\theta}))^2)_{i:n} = \sum_{i=1}^{h}((y_i - \boldsymbol{x}_i\boldsymbol{\theta})^2)_{i:n}.$$

□

Teorem 4.1 *Ako je T procjenitelj koji je regresijski ekvivarijantan, tada je*

$$\epsilon^*_n(T, Z) \leq \frac{\lfloor (n-p)/2 \rfloor + 1}{n},$$

za sve uzorke Z.

Dokaz: Vidi [10, str. 125].

Teorem 4.2 *Prijelomna vrijednost LTS procjenitelja s* $h = \frac{n}{2} + \frac{p+1}{2}$ *jednaka je*

$$\epsilon_n^*(T, Z) = \frac{\lfloor (n-p)/2 \rfloor + 1}{n}.$$

Dokaz: Vidi [10, str. 132].

Prema rezultatima prethodna dva teorema, za procjenitelja autokorelacijske funkcije $\widehat{\rho}_{LTS}$ najveća moguća asimptotska prijelomna vrijednost je 50%. Drugim riječima, kada bi uzorak bio kontaminiran više od 50% tada ne bi bilo moguće razlikovati dobre i loše podatke.

Napomena 4.1 *Kada promatramo vrijednost h za koju će prijelomna vrijednost biti najveća uočimo sljedeće. Ako s* $|A|$ *označimo broj dobrih opservacija, broj loših opservacija (stršećih vrijednosti)* $n - |A|$ *mora biti strogo manji od h, a s druge strane broj* $|A| + p + 1$ *mora biti najmanje h. Određivanje najbolje vrijednost h se stoga svodi na minimiziranje* $|A|$ *preko h, što dovodi do* $h = \frac{n}{2} + \frac{p+1}{2}$.

Napomena 4.2 *Može se pokazati da je asimptotska efikasnost LTS procjenitelja vrlo niska, iznosi oko 7%. Vidi [10, Poglavlje 3.4].*

4.3 MM procjenitelj

Svrha uvođenja MM procjenitelja je postizanje visoke prijelomne vrijednosti kao i visoke efikasnosti (uz uvjet da su greške regresije normalno distribuirane). Promotrimo sada S procjenitelje i M procjenitelje koje koristimo pri definiranju MM procjenitelja.
M procjenitelji su temeljeni na zamjeni kvadriranih reziduala s nekom drugom funkcijom reziduala, to jest

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \tau(r_i),$$

τ je neka simetrična funkcija s jedinstvenim minimumom u 0. Diferenciranjem prethodnog izraza s obzirom na θ_j dobivamo

$$\sum_{i=1}^{n} \psi(r_i)\boldsymbol{x_i} = \mathbf{0},$$

gdje je $\psi(r) = \tau'(r)$.
Navedeni LMS i LTS procjenitelji definirani su minimiziranjem robusne mjere rasipanja reziduala, a generaliziranjem postojećih ideja, Rousseeuw i Yohai[5] predstavljaju S procjenitelje. Za

[5]Victor Jaime Yohai (rođen 1939. g.), argentinski statističar

svaki vektor $\boldsymbol{\theta}$ možemo promatrati reziduale $r_i(\boldsymbol{\theta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\theta}$. S procjenitelj $\widehat{\boldsymbol{\theta}}$ je definiran s

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} s(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta})),$$

pri čemu je skala reziduala $\mathrm{s}(r_1(\boldsymbol{\theta}), \ldots, r_n(\boldsymbol{\theta}))$ definirana kao rješenje

$$\frac{1}{n}\sum_{i=1}^{n} \tau\left(\frac{r_i}{s}\right) = b, \tag{4.2}$$

b je obično jednak $E_\phi[\tau]$, gdje ϕ označava distribuciju standardne normalne slučajne varijable. Funkcija τ mora zadovoljavati sljedeće uvjete:
(S1) τ je simetrična, neprekidno diferencijabilna i vrijedi $\tau(0) = 0$,
(S2) postoji $c > 0$ takav da je τ strogo rastuća na $[\,0, c]$ i konstanta na $[\,c, \infty\rangle$.

Zbog uvjeta (S2), $\psi(x) = \tau'(x)$ će uvijek biti 0 nakon neke vrijednosti x, stoga je ψ padajuća funkcija. Ako postoji više od jednog rješenja jednakosti (4.2), tada je

$$\mathrm{s}(r_1, \ldots, r_n) = sup\{s; \frac{1}{n}\sum \tau(r_i/s) = b\}.$$

Ako ne postoji rješenje, tada je $\mathrm{s}(r_1, \ldots, r_n) = 0$.
MM procjenitelji se temelje na minimizaciji robusne M procjene skale reziduala. Definirani su u tri faze. U prvoj fazi određujemo nekog procjenitelja $\boldsymbol{T}_0$ s visokom prijelomnom vrijednosti, takvi procjenitelji su primjerice LMS ili LTS procjenitelji. U tu svrhu procjenitelj ne mora biti efikasan. Zatim određujemo M procjenu skale $s_n = s(\boldsymbol{r}(\boldsymbol{T}_0))$ definiranu izrazom (4.2) koristeći funkciju τ_0 koja zadovoljava (S1) i (S2). Neka je sada τ_1 neka druga funkcija koja također zadovoljava (S1) i (S2) i konačno MM procjenitelj $\boldsymbol{T}_1$ je definiran kao rješenje

$$\sum_{i=1}^{n} \psi_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right) \boldsymbol{x_i} = \mathbf{0}, \tag{4.3}$$

koje zadovoljava

$$S\left(\boldsymbol{T}_1\right) \leq S\left(\boldsymbol{T}_0\right),$$

gdje je

$$S\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \tau_1\left(\frac{r_i(\boldsymbol{\theta})}{s_n}\right),$$

pri čemu je $\tau_1(\frac{0}{0})$ definirano kao 0.

Napomena 4.3 *Zbog načina na koji je definiran MM procjenitelj, specijalno zbog prve faze u kojoj uzimamo nekog procjenitelja $\boldsymbol{T}_0$ s visokom prijelomnom vrijednosti, za očekivati je da će*

MM procjenitelj naslijediti neka svojstva procjenitelja $\boldsymbol{T}_0$. Vrijedi sljedeće: ako je procjenitelj $\boldsymbol{T}_0$ regresijski i afino ekvivarijantan, tada je i MM procjenitelj regresijski i afino ekvivarijantan. Vidi [11].

Prema prethodnoj napomeni i *Teoremu 4.1*, najveća asimptotska prijelomna vrijednost procjenitelja autokorelacijske funkcije $\widehat{\rho}_{MM}$ je 50%.

Razmotrimo sada asimptotsku normalnost MM procjenitelja u cilju određivanja asimptotske efikasnosti. Prije navođenja teorema koji govori o tome, uvedimo neke uvjete.
U klasičnom regresijskom modelu $y_i = \boldsymbol{\theta}_0' \boldsymbol{x}_i + e_i$, $i = 1, \ldots, n$ neka je $\boldsymbol{z}_i = (y_i, \boldsymbol{x_i})$ $(p+1)$-dimenzionalan vektor, gdje je $\boldsymbol{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$. Neka je $G_0(\boldsymbol{x})$ funkcija distribucije slučajnog vektora $\boldsymbol{X} = (X_1, \ldots, X_p)$ kojim modeliramo podatke $\boldsymbol{x}_i$ te $F_0(e)$ funkcija distribucije slučajne varijable ϵ kojom modeliramo e_i. Tada je funkcija distribucije slučajnog vektora $\boldsymbol{Z}$ kojim modeliramo $\boldsymbol{z}_i$ dana s

$$H_0(\boldsymbol{z}) = G_0(\boldsymbol{x}) F_0(y - \boldsymbol{\theta}_0' \boldsymbol{x}). \tag{4.4}$$

Dodatno pretpostavimo da G_0 ima drugi moment i da je

$$\boldsymbol{V} = E_{G_0}(\boldsymbol{x}_i \boldsymbol{x}_i')$$

regularna. Idući teorem govori o asimptotskoj normalnosti M procjenitelja, to uključuje i poseban slučaj MM procjenitelja.

Teorem 4.3 *Neka $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ dolaze iz distribucije H_0. Neka je τ_1 funkcija koja zadovoljava uvjete (S1) i (S2) te neka distribucija H_0 zadovoljava (4.4). Nadalje, neka je s_n skala reziduala koja jako konvergira prema σ_0 i $\boldsymbol{T}_n$ niz procjenitelja koji zadovoljavaju (4.3) i konzistentni su za vrijednost parametra $\boldsymbol{\theta}_0$. Tada vrijedi*

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma_0 \left[A(\psi_1, F_0)/B^2(\psi_1, F_0)\right] \boldsymbol{V}^{-1}\right),$$

gdje $\xrightarrow{d}$ označava konvergenciju po distribuciji i pri čemu je

$$A(\psi, F) = E_F(\psi^2(e/\sigma_0))$$

i

$$B(\psi, F) = E_F(\psi'(e/\sigma_0)).$$

Dokaz: Vidi [11, str. 42].

Prilikom definiranja MM procjenitelja, funkcije τ_0 i τ_1 možemo odabrati na sljedeći način. Neka je τ neka funkcija koja zadovoljava uvjete (S1) i (S2) i neka je $0 < k_0 < k_1$. Neka je $\tau_0(e) = \tau(e/k_0)$ i $\tau_1(e) = \tau(e/k_1)$.
Pretpostavimo da je MM procjenitelj dobiven koristeći u drugoj fazi s_n s vrijednosti $b = E_\phi(\tau(e/k_0))$. Odabirom odgovarajuće vrijednosti k_0 osiguravamo visoku prijelomnu vrijednost. Neka greške e_i imaju distribuciju $\mathcal{N}(0,1)$, tada je $\sigma_0 = 1$ i prema *Teoremu 4.3* asimptotska varijanca MM procjenitelja ovisi samo o k_1. Stoga, možemo odabrati k_1 tako da procjenitelj ima visoku efikasnost bez da narušimo prijelomnu vrijednost, koja ovisi samo o izboru k_0.
Na primjer, neka je funkcija τ zadana na sljedeći način

$$\tau(e) = \begin{cases} e^2/2 - e^4/2 + e^6/6 & \text{, za} \quad |e| \leq 1 \\ 1/6 & \text{, za} \quad |e| > 1. \end{cases}$$

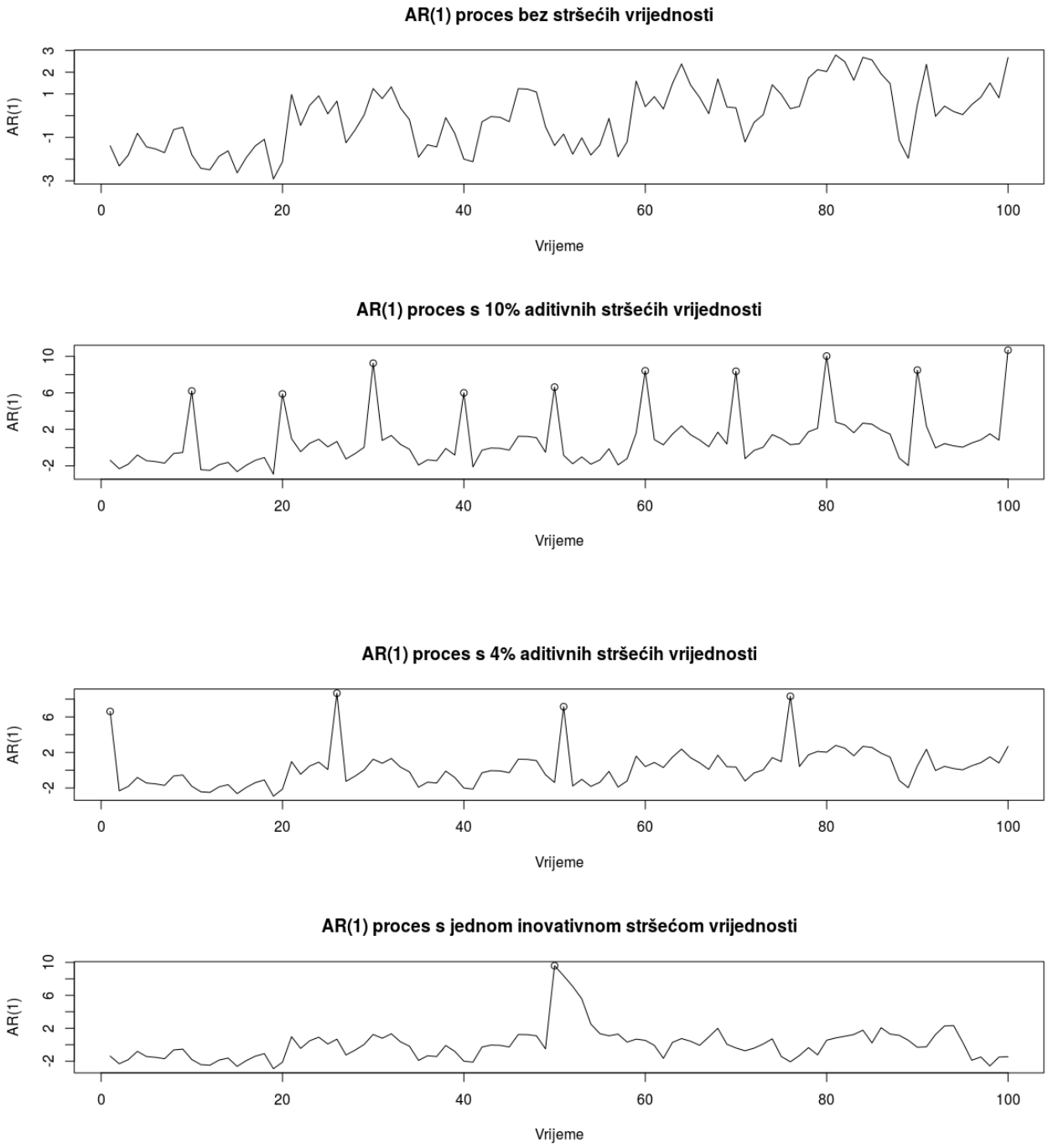
Ako je k_0 1.56 i pripadajuća vrijednost b je 0.0833 to bi osiguravalo visoku prijelomnu vrijednost, a vrijednost k_1 koja daje visoku efikasnost 0.95, za normalno distribuirane greške, je 4.68.

5 Usporedba procjenitelja na primjerima

Nakon što smo opisali klasičnog i tri robusna procjenitelja autokorelacijske funkcije, njihovu usporedbu promatrat ćemo na primjerima simulacija AR(1) i MA(1) procesa. Simulirat ćemo procese bez i sa stršećim vrijednostima. Također, razlikovat ćemo dvije vrste stršećih vrijednosti o kojima govorimo u nastavku.
Aditivna stršeća vrijednost (engl. *additive outlier*) utječe samo na jedan podatak čija je vrijednost iznenađujuće velika ili mala. Naime, neka je $\{X_t\}$ slabo stacionaran proces koji promatramo, te neka je $\{V_t\}$ stacionaran proces za koji vrijedi $P(V_t = 0) = 1 - \epsilon$ (u praksi ϵ je vrlo mali). Kada govorimo o procesu s aditivnim stršećim vrijednostima, umjesto $\{X_t\}$ zapravo promatramo $Y_t = X_t + V_t$, pri čemu su procesi $\{X_t\}$ i $\{V_t\}$ međusobno nezavisni. Aditivne stršeće vrijednosti opisane su kao velike pogreške u opažanju koje utječu samo na jedno promatranje. Inovativne stršeće vrijednosti (engl. *innovation outliers*) utječu ne samo na jedan podatak nego i na podatke nakon njega. Neka je zbog jednostavnosti $\{X_t\}$ AR(1) proces, $X_t = \phi X_{t-1} + Z_t$, $|\phi| < 1$, gdje je $\{Z_t\}$ IID šum s očekivanjem 0 i konačnom varijancom. Inovativna stršeća vrijednost je stršeća vrijednost procesa $\{Z_t\}$, stoga možemo reći da proces ima inovativne stršeće vrijednosti kada je distribucija procesa $\{Z_t\}$ s teškim repovima (na primjer Studentova t-distribucija).

Primjer 5.1 Simulirajmo AR(1) proces s koeficijentom $\phi = 0.8$, veličine uzorka $n = 100$ i $n = 300$. Promatrat ćemo proces bez stršećih vrijednost, s 10% i s 4 % aditivnih stršećih

vrijednosti te s jednom inovativnom stršećom vrijednosti. Na Slici 5.1 vidimo grafičke prikaze spomenutih simulacija za $n = 100$.



Slika 5.1: Simulacije AR(1) procesa $X_t = 0.8X_{t-1} + Z_t, t \in \mathbb{Z}$

Kako bi dobili bolji uvid u kvalitetu procjenitelja, napravit ćemo 50 simulacija trajektorija AR(1) procesa. Za svaku trajektoriju promatrat ćemo procjenu autokorelacijske funkcije

na koraku 1 i 2, zatim ćemo prosjek svih 50 procjena (u oznaci $\widehat{\rho}(1)$ i $\widehat{\rho}(2)$) usporediti sa stvarnim vrijednostima koje su $\rho(1) = \phi = 0.8$, $\rho(2) = \phi^2 = 0.64$. Kao mjeru kvalitete modela koristit ćemo srednje kvadratnu grešku (engl. *mean squared error*) koja je definirana izrazom $MSE = \frac{1}{n}\sum_{k=1}^{n}(\widehat{\theta}_k - \theta_k)^2$, pri čemu je $\widehat{\theta}_k$ procijenjena, a θ_k prava vrijednost u trenutku k.

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE	$\widehat{\rho}(2)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.7776534	0.006755597	0.5745129	0.06306935
		$\widehat{\rho}_{L1}$	0.7505075	0.01143263	0.5640363	0.07019928
		$\widehat{\rho}_{LTS}$	0.7589828	0.009146175	0.5751285	0.06646865
		$\widehat{\rho}_{MM}$	0.7631975	0.006988287	0.5750621	0.06284884
	300	$\widehat{\rho}_{OLS}$	0.7893631	0.00113256	0.6239262	0.03497851
		$\widehat{\rho}_{L1}$	0.7856752	0.001957156	0.6168119	0.03740505
		$\widehat{\rho}_{LTS}$	0.7890812	0.001518136	0.6204613	0.03609922
		$\widehat{\rho}_{MM}$	0.7895193	0.00111339	0.6209808	0.03533306

Tablica 5.1: Procjenitelji autokorelacijske funkcije AR(1) procesa bez stršećih vrijednosti

Prema Tablici 5.1, svi procjenitelji su podjednako dobri, što je bilo i za očekivati zbog toga sto se radi o simulaciji bez stršećih vrijednosti. Možemo primijetiti da $\widehat{\rho}_{OLS}$ nije bitno lošiji od robusnih procjenitelja. Povećanjem uzorka procjena postaje bolja te svi procjenitelji imaju vrlo malu srednje kvadratnu grešku.

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE	$\widehat{\rho}(2)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.1723605	0.4000033	0.1096671	0.4820832
		$\widehat{\rho}_{L1}$	0.2769194	0.2902508	0.1644626	0.4117584
		$\widehat{\rho}_{LTS}$	0.6790973	0.06878512	0.3473155	0.2592485
		$\widehat{\rho}_{MM}$	0.3654362	0.247537	0.173555	0.3975378
	300	$\widehat{\rho}_{OLS}$	0.174569	0.3934006	0.127415	0.4549351
		$\widehat{\rho}_{L1}$	0.2807523	0.2748059	0.1884689	0.3781115
		$\widehat{\rho}_{LTS}$	0.7804114	0.003013751	0.4341683	0.1660342
		$\widehat{\rho}_{MM}$	0.3272564	0.2580268	0.1848441	0.3806677

Tablica 5.2: Procjenitelji autokorelacijske funkcije AR(1) procesa s 10% aditivnih stršećih vrijednosti

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE	$\widehat{\rho}(2)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.3555732	0.2106867	0.2770685	0.288178
		$\widehat{\rho}_{L1}$	0.5787205	0.06189096	0.3395256	0.234481
		$\widehat{\rho}_{LTS}$	0.7435623	0.009591864	0.4792419	0.1427383
		$\widehat{\rho}_{MM}$	0.6933385	0.03688803	0.3544028	0.2224957
	300	$\widehat{\rho}_{OLS}$	0.389119	0.1717885	0.315768	0.2396096
		$\widehat{\rho}_{L1}$	0.6138811	0.03861774	0.4044656	0.1645691
		$\widehat{\rho}_{LTS}$	0.7871986	0.001202032	0.5778227	0.06127272
		$\widehat{\rho}_{MM}$	0.7835053	0.001544149	0.3859499	0.1814995

Tablica 5.3: Procjenitelji autokorelacijske funkcije AR(1) procesa s 4% aditivnih stršećih vrijednosti

Procjena autokorelacijske funkcije u slučaju kada imamo 10% aditivnih stršećih vrijednosti prikazana je u Tablici 5.2. Na ovom primjeru jasno vidimo kako veći broj stršećih vrijednosti predstavlja veliki problem za procjenu. Kako smo ranije naveli, asimptotska prijelomna vrijednost procjenitelja $\widehat{\rho}_{OLS}$ i $\widehat{\rho}_{L1}$ je 0% te zbog toga oba procjenitelja u ovom slučaju imaju vrlo loše rezultate. Iako procjenitelj $\widehat{\rho}_{MM}$ ima dobra svojstva, visoku prijelomnu vrijednost i visoku asimptotsku efikasnost, ovdje on nije bolji nego klasični procjenitelj. Jednako loše rezultate daje i s povećanjem uzorka. Procjenitelj $\widehat{\rho}_{LTS}$ za razliku od preostala tri procjenitelj ima zadovoljavajuće dobru procjenu i malu srednje kvadratnu grešku. Na koraku 2 procjena je nešto lošija no povećanjem uzorka dobivamo bitno manju srednje kvadratnu grešku. Možemo reći kako je procjenitelj temeljen na LTS metodi najpouzdaniji kada imamo veliki broj stršećih vrijednost.

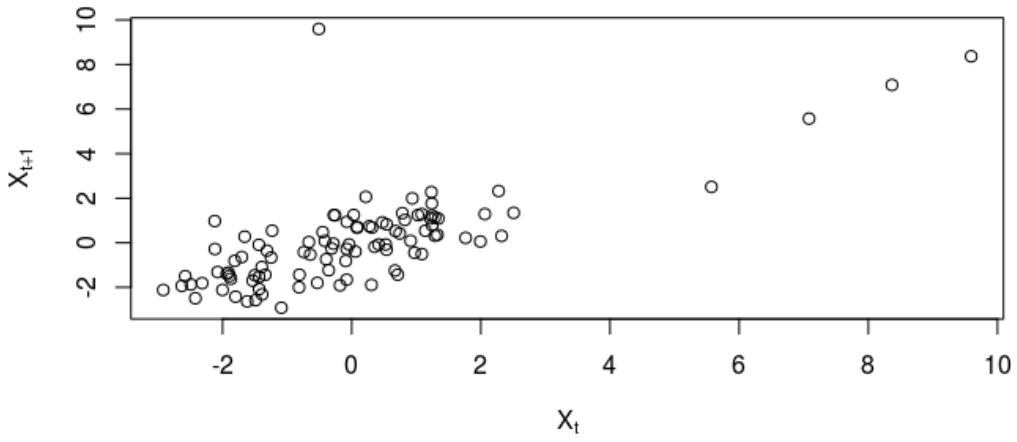
Razlika klasičnog i robusnih procjenitelja je jasno vidljiva i na primjeru AR(1) procesa s 4% aditivnih stršećih vrijednosti što nam prikazuje Tablica 5.3. Srednje kvadratna greška klasičnog procjenitelja je zamjetno veća u usporedbi s preostalim procjeniteljima. Od robusnih procjenitelja najlošije rezultate opet daje $\widehat{\rho}_{L1}$ no u ovom slučaju kada imamo manje stršećih vrijednost procjena je ipak bitno bolja. Procjenitelj temeljeni na MM i LTS regresijskoj metodi imaju podjednako male srednje kvadratne greške za procjenu na koraku 1 te su zbog toga najpouzdaniji za ovakvu vrstu procjene.

Prema Tablici 5.4, kada promatramo procjenu autokorelacijske funkcije za AR(1) proces s jednom inovativnom stršećom vrijednosti, svi procjenitelji imaju dobre rezultate. Razlog tome je što inovativna stršeća vrijednost AR(1) procesa izvlači samo jednu točku izvan regresi-

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE	$\widehat{\rho}(2)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.7576742	0.006077746	0.604899	0.04728638
		$\widehat{\rho}_{L1}$	0.7843233	0.003572742	0.6167488	0.04139327
		$\widehat{\rho}_{LTS}$	0.7789106	0.004437371	0.6155922	0.04253858
		$\widehat{\rho}_{MM}$	0.7892352	0.002058294	0.6178625	0.03889516
	300	$\widehat{\rho}_{OLS}$	0.795174	0.001056892	0.6296862	0.03284892
		$\widehat{\rho}_{L1}$	0.7948611	0.0013609	0.6313293	0.03161406
		$\widehat{\rho}_{LTS}$	0.7938963	0.001682056	0.641898	0.02771934
		$\widehat{\rho}_{MM}$	0.7917896	0.001326918	0.6274165	0.03332303

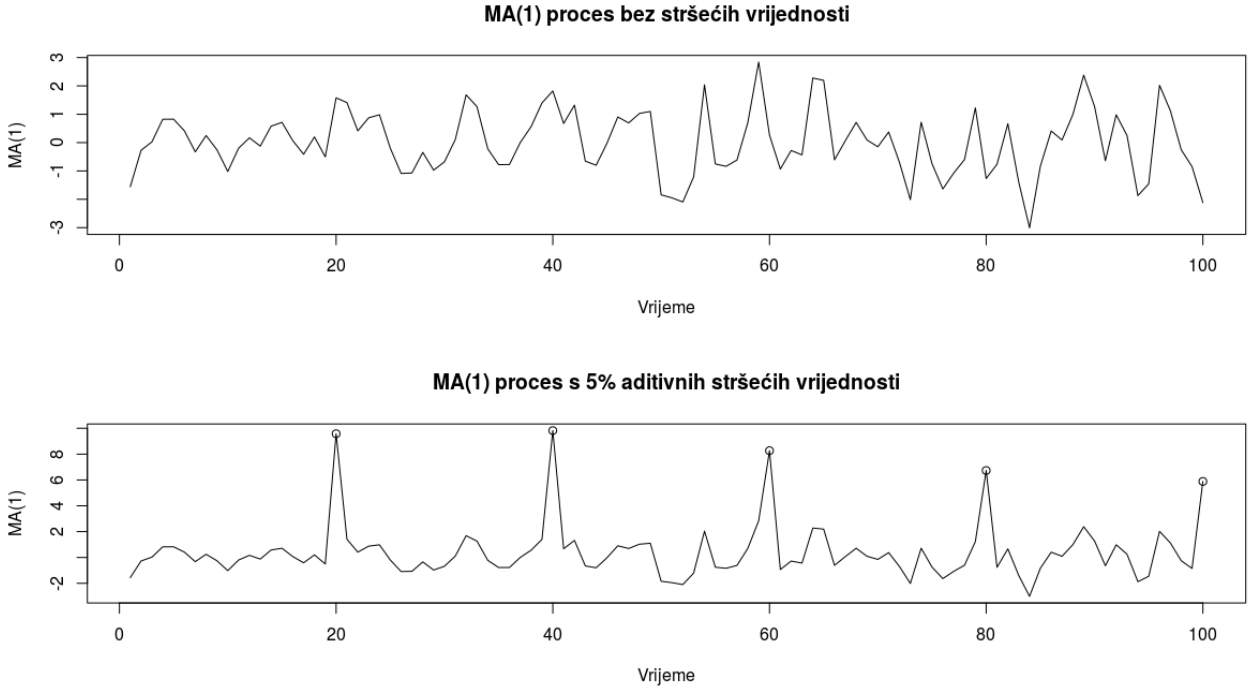
Tablica 5.4: Procjenitelji autokorelacijske funkcije AR(1) procesa s jednom inovativnom stršećom vrijednosti

jske linije dok druge ekstremne točke generira na regresijsku liniju. Na to ukazuje Slika 5.2. U ovom slučaju nema značajnije razlike klasičnog procjenitelja u usporedbi s robusnim.



Slika 5.2: Dijagram raspršenja (X_t, X_{t+1}) AR(1) procesa s jednom inovativnom stršećom vrijednosti

Primjer 5.2 U ovom primjeru promatrat ćemo MA(1) proces $X_t = Z_t + 0.8Z_{t-1}, t \in \mathbb{Z}$, napravit ćemo simulacije procesa bez stršećih vrijednosti te s 5% aditivnih stršećih vrijednosti. Sljedeća slika prikazuje grafičke prikaze simulacija procesa za veličinu uzorka $n = 100$.



Slika 5.3: Simulacije MA(1) procesa $X_t = Z_t + 0.8Z_{t-1}, t \in \mathbb{Z}$

Napravit ćemo 50 simulacija trajektorija MA(1) procesa i promatrati procjenu autokorelacijske funkcije na koraku 1, prosjek svih 50 procjena usporedit ćemo sa stvarnom vrijednosti autokorelacijske funkcije koja je $\rho(1) = \frac{\theta}{1+\theta^2} \approx 0.4878049$. Prema *Teoremu 1.1* MA(1) proces je 1-koreliran, odnosno $\rho(h) = 0$ za $h > 1$. Zbog toga procjenu ne promatramo za korake veće od 1. Kao i u prošlom primjeru, za mjeru kvalitete modela koristit ćemo srednje kvadratnu grešku.

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.4649924	0.006138543
		$\widehat{\rho}_{L1}$	0.4831999	0.0115285
		$\widehat{\rho}_{LTS}$	0.4577198	0.01204496
		$\widehat{\rho}_{MM}$	0.4634712	0.006997747
	300	$\widehat{\rho}_{OLS}$	0.4830232	0.001528828
		$\widehat{\rho}_{L1}$	0.4838641	0.003121628
		$\widehat{\rho}_{LTS}$	0.4814148	0.00237314
		$\widehat{\rho}_{MM}$	0.4814597	0.001606178

Tablica 5.5: Procjenitelji autokorelacijske funkcije MA(1) procesa bez stršećih vrijednosti

ϕ	n	procjenitelj	$\widehat{\rho}(1)$	MSE
0.8	100	$\widehat{\rho}_{OLS}$	0.1549919	0.1154746
		$\widehat{\rho}_{L1}$	0.2049782	0.08946825
		$\widehat{\rho}_{LTS}$	0.3263723	0.05044069
		$\widehat{\rho}_{MM}$	0.1925337	0.09253538
	300	$\widehat{\rho}_{OLS}$	0.140442	0.1219756
		$\widehat{\rho}_{L1}$	0.1808763	0.096473
		$\widehat{\rho}_{LTS}$	0.360274	0.03724465
		$\widehat{\rho}_{MM}$	0.1734749	0.1001008

Tablica 5.6: Procjenitelji autokorelacijske funkcije MA(1) procesa s 5 % aditivnih stršećih vrijednosti

Prema Tablici 5.5 svi procjenitelji su podjednako dobri za procjenu autokorelacijske funkcije MA(1) procesa. Kako u ovom slučaju nisu prisutne stršeće vrijednosti, klasični procjenitelj $\widehat{\rho}_{OLS}$ je najpouzdanijeg zbog najmanje srednje kvadratne greške. Procjena se s povećanjem uzorka očekivano poboljšava za sve procjenitelje.
Promatranjem Tablice 5.6 koja prikazuje procjenu autokorelacijske funkcije MA(1) procesa s 5% aditivnih stršećih vrijednosti dolazimo do sličnih zaključaka kao i u prethodnom primjeru s AR(1) procesom. Procjenitelj $\widehat{\rho}_{LTS}$ ima najbolje rezultate u usporedbi s ostalim procjeniteljima. Procjenitelji $\widehat{\rho}_{OLS}$ i $\widehat{\rho}_{L1}$ zbog male prijelomne vrijednosti nisu pouzdani, no čak i bolja teorijska svojstva procjenitelja $\widehat{\rho}_{MM}$ nisu rezultirala dobrim rezultatima tog procjenitelja. Iako smo kao svojstvo procjenitelja $\widehat{\rho}_{LTS}$ istaknuli vrlo malu asimptotsku efikasnost, to u ovim primjerima nije bilo vidljivo.

Literatura

[1] P. J. BROCKWELL, R. A. DAVIS, *Introduction to Time Series and Forecasting*, Springer, New York, 2002.

[2] P. J. BROCKWELL, R. A. DAVIS, *Time Series: Theory and Methods* ,Springer, New York, 1991.

[3] C. C. CHANG, D. N. POLITIS, *Robust Autocorrelation Estimation*, Journal of Computational and Graphical Statistics, Vol. 25 (2016.), 144-166

[4] H. HASSANI , *Sum of the sample autocorrelation function*, Random Operators and Stochastic Equations 17 (2009), 125-130

[5] J. JUREČKOVÁ, J. PICEK, M. SCHINDLER, *Robust Statistical Methods with R, Second Edition*,Chapman and Hall/CRC, 2019.

[6] J.-P. KREISS ,E. PAPARODITIS, D.N. POLITIS, *On the Range of Validity of the Autoregressive Sieve Bootstrap*, The Annals of Statistics, Vol. 39, No. 4 (2011.), 2103–2130

[7] R.A. MARONNA, R.D. MARTIN, V.J. YOHAI, M. SALIBIÁN-BARRERA, *Robust Statistics: Theory and Methods (with R)*, John Wiley & Sons, 2019.

[8] T.C MILLS , *The Econometric Modelling of Financial Time Series* , Cambridge University Press, 1999.

[9] K. NEUSSER, *Time Series Econometrics*, Springer, New York,2016.

[10] P. ROUSSEEUW, A. LEROY, *Robust Regression and Outlier Detection*,New Jersey, Wiley-Interscience, 1987.

[11] V.J. YOHAI, *High Breakdown-Point and High Efficiency Robust Estimates for Regression*, Technical report, Department of Statistics, University of Washington (1985.)

Sažetak

U ovom diplomskom radu obrađena je procjena autokorelacijske funkcije. Na samom početku rada napravljena je teorijska osnova vremenskih nizova, uveden je pojam stacionarnosti i opisani su osnovni modeli stacionarnih vremenskih nizova. Zatim, objašnjena je procjena očekivanja koja je bitna karakteristika stacionarnih procesa. U nastavku rada navedeno je nekoliko vrsta procjenitelja autokorelacijske funkcije. Kao klasični procjenitelj promatrana je uzoračka funkcija autokorelacija. Kako postojanje stršećih vrijednosti među podacima značajno utječe na kvalitetu procjene uvedeni su robusni procjenitelji. Navedena su svojstva i prednosti procjenitelja dobivenih na temelju tri regresijske metode: $L1$ metoda, metoda najmanjih odrezanih kvadrata i MM metoda. Na kraju rada napravljena je usporedba procjenitelja na simulacijama stacionarnih procesa koja ukazuje na prednosti korištenja robusnih procjenitelja.

Ključne riječi: vremenski niz, stacionarnost, očekivanje, autokorelacijska funkcija, uzoračka autokorelacijska funkcija, linearna regresija, robusnost

Summary

This diploma thesis presents an estimation of autocorrelation function. In the first part, theoretical basis of time series, concept of stationarity and the most important models of stationary time series are described. Furthermore, an estimate of the mean is explained, which is an important characteristic of stationary processes. Several types of estimators of autocorrelation function are presented. Sample autocorrelation function was observed as a classical estimator. Since the presence of outliers among data can significantly affect the quality of estimation, robust estimators have been introduced. The properties of estimators based on three regression methods are given: $L1$ method, least trimmed squares and MM method. At the end of the paper, a comparison of estimators on stationary process simulations is made, which indicates the advantages of using robust estimators.

Key words: time series, stationarity, mean, autocorrelation function, sample autocorrelation function, linear regression, robustness

Životopis

Rođena sam 24.11.1994. godine u Požegi u Republici Hrvatskoj. Obrazovanje sam započela u Osnovnoj školi Antun Kanižlić u Požegi nakon čega upisujem Opću gimnaziju u Požegi. Završetkom srednjoškolskog školovanja, 2013. godine postajem studentica Prirodoslovno-matematičkog fakulteta u Zagrebu gdje sam stekla naziv sveučilišne prvostupnice edukacije matematike. Obrazovanje nastavljam 2017. godine na Sveučilištu Josip Juraj Strossmayer u Osijeku upisom Diplomskog studija matematike, smjer Financijska matematika i statistika. Tijekom studija odradila sam stručnu praksu u Erste & Steiermärkische Banci u Osijeku.