Univerzitet u Novom Sadu
Prirodno-matematički fakultet
Departman za matematiku i informatiku

Tomišić Mihailo

Robusne statistike i njihove primene u teoriji kredibiliteta

Master rad

Mentor:
Prof. dr Dora Seleši

2017, Novi Sad

Sadržaj

Glava 1

Robusne metode

1.1 Uvod

Statistička analiza podataka zavisi od prirode podataka koji se analiziraju, kao i od metode kojom se analiziraju ovi podaci. Metoda je način analize podataka. Optimalan algoritam analize podataka zavisi od odabira metode i prirode podataka koji se analiziraju, odnosno implicitnih ili eksplicitnih pretpostavki o samim podacima. To na primer mogu biti pretpostavke o slučajnosti, nezavisnosti populacije, o raspodeli populacije, o raspodeli parametra koji se ocenjuju, itd. Ovo su, dakle, pretpostavke na koje se model oslanja. Neke od ovih pretpostavki su matematički pogodne racionalizacije i retko kada su u potpunosti tačne na uzorcima.

Optimalan izbor metoda zavisi od dostupnih informacija o podacima koji se analiziraju. U fokusu ovog rada su isključivo tačkaste ocene parametara, preciznje reč je o jednoj klasi robusnih tačkastih ocena parametara.

U odnosu na *a priori* dostupne informacije o populaciji podataka optimalan izbor metoda dat je Tabelom 1.1.

Tabela 1.1: Optimalan izbor metoda

Slučajnost i dostupne informacije	Metod
$f(x;\theta)$, slučajno θ	Bajezove metode
$f(x;\theta)$, nepoznato θ	Parametarski pristup: *MLE*
$f(x;\theta) \in \mathscr{F}$ ($\mathscr{F}$ je okolina simetrične unimodalne raspodele)	Robusne metode
$f(x;\theta) \in \mathscr{F}$ ($\mathscr{F}$ je uopštena klasa raspodela)	Neparametarske metode

Robusnost je osobina predmeta ili osobina sistema koja predstavlja otpornost u smislu neosetljivosti na male promene (devijaciju), odnosno, neosetljivost na neki stres koji deluje na posmatrani subjekat ili generalnu neosetljivost na greške.

Sa druge strane, fragilnost je osobina predmeta ili osobina sistema koja opisuje osetljivost na promene, stres koji utiče na posmatrani subjekat.

Klasične, odnosno nerobusne metode su upravo osetljive na raspodele sa dugačkim repom i osetljive su na male promene u vrednostima podataka. Njihova osetljivost se može manifestovati autlajerima[1] i tačkama leveridža (jakog uticaja)[2] populacije koja se analizira. Prirodno se postavlja pitanje detekcije i odbacivanja ovih vredosti. U literaturi se mogu naći primeri gde samo 2% populacije podataka pogoršava model odnosno ocenjivač i vrše jako veliku inflaciju relevantnih statističkih pokazatelja tog ocenjivača.

Odbacivanje autlajera je nekada u potpunosti opravdano i poželjno, dok u nekim drugim slučajevima ovaj proces samo po sebi predstavlja problem. Ova problematika odnosi se na detekciju autlajera i ovaj problem nije uvek jednostavan i adekvatno rešiv u smislu jasne kategorizacije ovih vrednosti. Prve diskusije o podesnosti odbacivanju autlajera datiraju još od Danijela Bernulija (1877), Basela i Bayera (1838). Student (1927) predlaže dodatne observacije u slučaju pojave autlajera sa kombinacijom odbacivanja. Empirijski je pokazano da u nekim slučajevima metode bazirane na odbacivanju autlajera nikada u potpunosti ne dostižu učinak robusnih metoda. Videti Hampel (1974a, 1976). Takođe pretpostavka o nezavisnosti uzorka koji se analizira ne važi uvek u potpunosti, već se često pojavljuje mala zavisnost među podacima. Kod modela regresije najčešće se javlja kolinearnost, odnosno uzajamna zavisnost u objašnjavajućim promenljivima i ovoj pojavi neće biti posvećena pažnja.

Robusne metode se pojavljuju kao prirodna ekstenzija klasičnih metoda iz potrebe za grubljim metodama koje su neosetljive na pomenute devijacije u pretpostavkama modela, bilo da je reč o raspodeli ili neosetljivosti na promene vrednosti podataka. Cilj rada je diskusija o robusnim osobinama statistika i robusnim metodama koje otklanjaju probleme koje klasične (nerobusne) metode tačkastih ocena parametara ne uspevaju adekvatno da reše. Devijacije u podacima se često javljaju zbog zaokruživanja vrednosti i grešaka koje nastaju pri očitavanju registrovanih vrednosti, bilo da je to produkt ljudskog ili mašinskog faktora. Takođe prilikom pravljenja modela jako bitan uticaj igra okolina i usovi pod kojim je određeni događaj ispoljen.

Sama robusnost metoda se postiže odricanjem određenog dela efikasnosti, što nužno ne mora da znači da se dobijaju lošiji modeli, već to zavisi, kao što je već pomenuto, i od uzorka podataka koji se obrađuje. Robusne metode se u literaturi javljaju oko 1950-te godine i od tada su aktivno razvijane. Danas robusnost metode ima više značenja i različitih pristupa. Ovaj rad obrađuje dva pristupa koji se javljaju u literaturi negde oko 1960 godine, a to su Huberov minimaks pristup (kvantitativna robusnost) i Hampelov pristup baziran na funkcijama uticaja (kvalitativna robusnost). Huber (1964) objavljuje rad pod nazivom *Robust estimation of location parametar* i oblikuje osnove teorije za robusno ocenivanje. U ovom radu on uvodi *M*-ocenjivače koji su generalizacija *ML* ocenjivača[3] i uvodi adekvatne pokazatelje

[1] Od engleske reči *outlier* - observacija koja je van očekivanog opsega vrednosti studije ili eksperimenta.

[2] Leverage points - su observacije koje imaju jak uticaj na model. Nužno se ne nalaze van populacije.

[3] Maximum Likelihood Estimation - Metoda Maksimalne Verodostojnoosti.

kojim se analizira robusnost ocenjivača. On pretpostavlja da model može da sadrži deo podataka koji ili dolazi iz neke nepoznate raspodele ili je produkt nekakve greške. Huberov cilj je da optimizuje najgori mogući scenario iz okoline modela[4] što je merio asimptotskom varijansom i asimptotskom pristrasnošću ocenjivača. Sem ova dva pristupa u literaturi se može naći još pristupa problematici, ali većina njih nije imala neke bitne posledice.

Hampelov pristup je baziran na tri osnovna koncepta. To su kvalitativna robusnost, uticajna funkcija i tačka preloma statistike. Ovim pokazateljima će jasno biti ilustrovane neke od mana klasičnih (nerobusnih) metoda. Sam pristup baziran je na činjenici da veliki broj statistika zavisi od empirijske funkcije raspodele i stoga ih je moguće posmatrati kao funkcionele na polju verovatnoća. Posmatranje niza funkcionela omogućava definisanje neprekidnosti i izvoda. Kvalitativna robusnost je definisana kao uniformna neprekidnost raspodela statistike kako i se obim uzorka povećava usko je povezana sa neprekidnošću statistike koju posmatramo kao funkcionelu na slaboj topologiji. Tačka preloma direktno meri pouzdanost statistike.

Generalno od robusne metode se zahteva da bude kvalitativno robusna, da ima visoku tačku preloma i da ima nisku ukupnu osetljivost na greške. Optimalnom procedurom se smatra ona procedura koja pravi kompromis izmedju svih relevantnih aspekata, ne samo maksimalno optimalne u odnosu na jedan ili dva izabrana askpekta.

Danas postoji veliki broj robusnih metoda koje su neretko razvijane nezavisno jedna od drugih i uglavnom su prilagođene nekoj specifičnoj problematici. U literaturi se takođe mogu naći i metode koje su robusne u odnosu na neku osobinu u smislu smanjene osetljivosti na konkretnu osobinu ili robusne metode koje su prilagođene različitim obimima populacije ili metodi koji ne zavise od obima populacije koja se analizira. Cilj ovog rada nije obrada svih robusnih metoda, već uvod u teoriju robusne statistike i njena primena, stoga su obrađeni i prezentovani samo neki od odabranih modela. Pažnja je posvećena metodima koji su dati jednom interacijom sve dok nije navedeno drugačije.

Ukratko od robusne metode ćemo očekivati da bude efikasna (u smislu neke optimalnosti), da poseduje određeni nivo stabilnosti i da se prilikom pojave većih devijacija od modela nužno ne kvari u potpunosti. Može se reći i da je primarni cilj robusnih metoda obezbeđenje od ogromnih grešaka.

U ostatku poglavlja navodimo pretpostavke koje važe u daljem delu rada. Neka je F_θ funkcija raspodele na nekom prostoru gde je $\theta \in \Theta$ nepoznati parametar koji je potrebno oceniti, a Θ neki prostor parametara. Za nepoznati parametar potpuno ekvivalentno koristimo oznake θ i $T(F)$. Sem ako nije navedeno drugačije pretpostavlja se da je $\Theta \subseteq \mathbb{R}$ otvoren i konveksan skup. Za razliku od klasičnog ocenivanja parametara gde uzorak međusobno nezavisan i prati istu raspodelu $\{F_\theta, \theta \in \Theta\}$ što je idealizovana aproksimacija realnosti, robusno ocenivanja parametara se vrši na nekoj okolini ove raspodele što je daleko prirodnije. Upravo u ovom procesu se gubi određena količina efikasnosti.

[4] Pod okolinom modela smatramo devijaciju od pretpostavki modela.

Ocenjivač nepoznatog parametra raspodele biće definisan na dva načina. Ocenjivač parametra je statistika ili niz statistika definisan preko $T_n = T_n(X_1, ..., X_n)$ na uzorku slučajnih promenljivih obima n, odnosno $(X_1, ..., X_n)$. Realizacija prostog slučajnog uzorka biće obeležavana sa $(x_1, ..., x_n)$. Sa druge strane statistika je definisana i preko funkcionele T ili niza funkcionela, na nekom prostoru, vezom $T_n(x_1, ..., x_n) = T(F_n)$ gde je F_n empirijska funkcija raspodele definisana u narednom delu teksta. Domen funkcionele T_n može biti prostor empirijskih raspodela, ali mnogo ćešće se upravo za ovaj prostor uzima prostor svih mera verovatnoće $\mathcal{M}$. Idealno, slučajne promenljive $(X_1, ..., X_n)$ su *i.i.d.*[5], odnosno one čine jedan prost slučajan uzorak obima n. Dalje se pretpostavlja da postoji preslikavanje $F \mapsto R$ na skupu svih raspodela za koje je T definisano, tako da

$$T_n(X_1, ..., X_n) \xrightarrow{p} T(F_\theta), \text{ kada } n \to \infty.$$

Dirakova mera skoncentrisana u x je obeležena sa δ_x.

1.1.1 Osnovni pojmovi teorije verovatnoće i statistike

Definicija 1. Niz slučajnih promenljivih $X_1, X_2, ...$ konvergira u verovatnoći ka slučajnoj promenljivoj X ako za svako $\varepsilon > 0$,

$$P\{|X_n - X| \geq \varepsilon\} \to 0, \text{ kada } \ n \to \infty.$$

Pišemo $X_n \xrightarrow{p} X$.

Teorema 1. *Ako je $g(x)$, $x \in R$ neprekidna funkcija i ako $X_n \xrightarrow{p} X$ tada*

$$g(X_n) \xrightarrow{p} g(X).$$

Definicija 2. Niz slučajnih promenljivih $X_1, X_2, ...$ konvergira skoro sigurno ka slučajnoj promenljivoj X ako

$$P\{X_n \to X, \ n \to \infty\} = 1.$$

Pišemo $X_n \xrightarrow{s.s.} X$.

Teorema 2. *Ako niz slučajnih promenljivih $X_1, X_2, \ldots$ skoro sigurno konvergira ka slučajnoj promenljivoj X, kada $n \to \infty$ onda taj niz konvergira i u verovatnoći ka X kada $n \to \infty$.*

Definicija 3. Niz slučajnih promenljivih $X_1, X_2, \ldots$ konvergira u raspodeli ka slučajnoj promenljivoj X, kada $n \to \infty$, ako niz odgovarajućih funkcija raspodele $F_{X_1}(x), F_{X_2}(x), \ldots$ kompletno konvergira ka funkciji $F_X(x)$, (konvergira za svako $x \in R \cup \{-\infty, \infty\}$ za koje je $F_X(x)$ neprekidna funkcija). Pišemo $X_n \xrightarrow{r} X$.

Definicija 4. Statistika $T_n = T(F_n)$ je konzistentna ako konvergira u verovatnoći ka $T(F)$. Niz $\{T_n\}_{n \in N}$ konvergira u verovatnoći ka slučajnoj promenljivoj $T(F)$ ako $\forall \varepsilon > 0$

$$P\{\left|T(F_n) - T(F)\right| \geq \varepsilon\} \to 0, n \to \infty.$$

[5]independent, identically distributed.

Definicija 5. Asimptotska varijansa definisana je kao

$$Asimp.Var[\sqrt{n}(T(F_n) - T(F))] = \lim_{n\to\infty} E[\sqrt{n}(T(F_n) - T(F))]^2.$$

Definicija 6. Pristrasnost ocenjivača definisana je kao

$$Bias[T(F_n)] = E[T(F_n)] - T(F).$$

Definicija 7. Asimptotska pristrasnost (ukoliko postoji) je definisana kao

$$\lim_{n\to\infty} E[T_n - T(F)].$$

Definicija 8. Statistika T_n je asimptotski normalna ako

$$\mathcal{L}_F\{\sqrt{n}[T_n - T(F)]\} \to \mathcal{N}(0, V(F,T)),$$

gde je $V(F,T)$ asimptotska varijansa u nekom okruženju $\mathcal{P}_\varepsilon(F)$ raspodele F.

Okruženja $\mathcal{P}_\varepsilon(F)$ su definisana u narednom poglavlju, dok oznaku $\mathcal{L}_F$ koristimo za funkciju raspodele date statistike.

Definicija 9. Srednje kvadratna greška ocenjivača definisana je kao

$$MSE(T(F_n)) = E[T(F_n) - T(F)]^2.$$

Odnosno,

$$\begin{aligned} MSE(T(F_n)) &= E[T(F_n) - E[T(F_n)]]^2 + [Bias[T(F_n)]^2 \\ &= V[T(F_n)] + [Bias[T(F_n)]^2. \end{aligned}$$

Očigledno sledi da za nepristrasne statistike važi $MSE(T(F_n)) = V[T(F_n)]$.

Definicija 10. Ocenjivač je asimptotski nepristrasan ukoliko važi

$$\lim_{n\to\infty} E[T_n] = T(F).$$

Lema 1. *Ako je ocenjivač nepristrasan, onda je on i asimptotski nepristrasan.*

Definicija 11. Ocenjivač $T(F_n)$ nepoznatog parametra $T(F)$ je efikasan ako

1. $T(F_n)$ je nepristrasan.
2. $Var[T(F_n)] \leq Var[\tilde{T}(F_n)]$ gde je $\tilde{T}(F_n)$ bilo koji ocenjivač parametra θ.

Definicija 12. Nejednakost Rao-Kramera je data sa

$$Var[\theta_i] \geq I^{ii},$$

gde je I^{ii} dijagonalni element Fišerove matrice informacija definisana u 45.

Definicija 13. Ocenjivač je najbolji linearni nepristrasan ocenjivač[6] ako ispunjava

1. $T(F_n)$ je linearna funkcija opažanja iz uzorka.
2. $T(F_n)$ je nepristrasan.
3. $T(F_n)$ ima minimalnu varijansu u klasi svih linearnih nepristrasnih ocenjivača.

Ocenjivač nepoznatog parametra ne može biti efikasan (u smislu navedene definicije) ukoliko ne koristi sve dostupne informacije iz uzorka.

[6] *BLUE* - best linear unbiased estimator or *MVLUE* (skraćeno od minimum variance).

1.2 Slaba topologija i njena metrizacija

Definicija 14. Neka je X neprazan skup. Funkcija $d : X^2 \mapsto [0, +\infty)$ je metrika na skupu X ako i samo ako za sve $x, y, z \in X$ važi

$M1$ $d(x, y) \geq 0$ i $d(x, y) = 0$ ako i samo ako je $x = y$.

$M2$ $d(x, y) = d(y, x)$.

$M3$ $d(x, y) \leq d(x, z) + d(z, y)$.

Tada se par (X, d) naziva metrički prostor. Broj $d(x, y)$ nazivamo rastojanjem između x i y.

Definicija 15. Topološki prostor $(X, \mathcal{O})$ je metrizabilan ako i samo ako postoji metrika d na skupu X takva da je $\mathcal{O} = \mathcal{O}_d$. Tada kažemo da je topologija indukovana metrikom d.

Teorema 3. *Neka su (X, d_X) i (Y, d_Y) metrički prostori i $f : X \mapsto Y$. Preslikavanje f je neprekidno u tački $x_0 \in X$ ako i samo ako*

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall x \in X)(d_X(x, x_0) < \delta \Rightarrow d_Y(f(x), f(x_0)) < \varepsilon).$$

Definicija 16. Neka je (X, d) metrički prostor. Za niz $\{x_n : n \in N\}$ u prostoru X kažemo da je Košijev niz ako i samo ako važi

$$(\forall \varepsilon > 0)(\exists n_0 \in N)(\forall m, n \geq n_0)(d(x_m, x_n) < \varepsilon).$$

Teorema 4. *U proizvoljnom metričkom prostoru (X, d) važi:*

1. *Svaki Košijev niz je ograničen.*
2. *Ako Košijev niz ima konvergentan podniz, onda je i sam konvergentan.*
3. *Svaki konvergentan niz je Košijev.*

Teorema 5. *Neka je (X, d) metrički prostor, onda važi:*

1. *Ako je $\{x_n\}$ niz u X i $x \in X$, onda je $\lim x_n = x$ ako i samo ako važi*

$$(\forall \varepsilon > 0)(\exists n_0 \in N)(\forall n \geq n_0)(d(x_n, x) < \varepsilon).$$

2. *Ako postoji, granica niza je jedinstvena.*
3. *Svaki konvergentan niz je ograničen.*

Definicija 17. Metrički prostor (X, d) je kompletan ako i samo ako je svaki Košijev niz u X konvergentan.

Definicija 18. Neka je $(X, \mathcal{O})$ topološki prostor. Skup $D \subseteq X$ je gust u X ako i samo ako je $\overline{D} = X$[7]. Prostor $(X, \mathcal{O})$ je separabilan ako i samo ako postoji skup $D \subseteq X$ koji je gust i prebrojiv.

[7] $\overline{D}$ označava zatvaranje skupa D.

Definicija 19. Topološki prostor zadovoljava drugu aksiomu prebrojivosti ako i sako ako postoji baza $\mathcal{B}$ topologije $\mathcal{O}$ takva da je $card(\mathcal{B}) \leq \aleph_0$.

Teorema 6. *Ako topološki prostor $(X, \mathcal{O})$ zadovoljava drugu aksiomu prebrojivosti, onda je separabilan.*

Primer 1. *Familija svih otvorenih intervala $B = \{(a, b) : a, b \in \mathbb{R} \wedge a < b\}$ u skupu realnih brojeva je baza uobičajene topologije na skupu $\mathbb{R}$. Obeležava se sa $(\mathbb{R}, \mathcal{O}_{uob})$. Metrika koja određuje (metrizuje) ovaj prostor je data sa $d(x, y) = |x - y|$. Metrički prostor $(\mathbb{R}, \mathcal{O}_{uob})$ je kompletan i separabilan.*

Generalno oznakom Ω obeležava se uzorački porostor $\mathbb{R}^n$ konačnih dimenzija. U ovom poglavlju pretpostavlja se da je Ω poljski prostor.

Definicija 20. Topološki prostor $(X, \mathcal{O})$ je poljski prostor (Polish space) ukoliko je separabilan i metrizibilan metrikom d takvom da je (X, d) kompletan metrički prostor.

Primer 2. *Neki bitni poljski prostori su $(\mathbb{R}, \mathcal{O}_{uob})$, svaki separabilan Banahov prostor, Kantorov skup, otvoren interval $(0, 1), \ldots$*

Neka je $\mathcal{M}$ prostor svih mera verovatnoće na $(\Omega, \mathcal{B})$ gde je $\mathcal{B}$ Borelova σ-algebra (tj. najmanja σ algebra takva da sadrži sve otvorene podskupove Ω). Sa $\mathcal{M}'$ obeležava se skup konačnih mera sa predznakom (konačnih naboja) na $(\Omega, \mathcal{B})$, odnosno to je linearni prostor generisan sa $\mathcal{M}$. U slučaju $\Omega = \mathbb{R}$ biće korišćeno veliko slovo F i za meru i za pridruženu funkciju raspodele sa konvencijom da $F(\cdot) = F\{(-\infty, x)\}$, gde je $F\{\ \}$ skupovna funkcija.

Teorema 7. *Svaka konačna Borelova mera $F \in \mathcal{M}$ je regularna u smislu da svaki Borelov skup $B \in \mathcal{B}$ može biti aproksimirana u F-meri kompaktnim skupom C od dole i otvorenim skupom G od gore tako da*

$$\sup_{C \subset B} F\{C\} = F\{B\} = \inf_{G \supset B} F\{G\}. \tag{1.1}$$

Definicija 21. Slaba(-zvezdasta) topologija na $\mathcal{M}$ je najslabija topologija takva da za svaku ograničenu i neprekidnu funkciju ψ preslikavanje $T_F : \mathcal{M} \mapsto R$ dato sa

$$F \mapsto \int \psi dF,$$

odnosno, preciznije zapisano

$$F(\psi) = \int \psi dF = \langle F, \psi \rangle,$$

je neprekidno.

Iz navedene definicije intutivno se vidi da ψ mora biti neprekidna i ograničena funkcija, inače jedna mala strateški odabrana pomerena vrednost uzorka $(x_1, \ldots, x_n)$ ili promena nastala dodavanjem mase δ_x funkciji F_n može naneti ogromne promene

u vrednosti statistike $T(F_n) = \int \psi dF_n$.

Neka je L linearna funkcionela na $\mathcal{M}$ (ili, preciznije restrikcija na $\mathcal{M}$ linearnih funkcionela definisanih na $\mathcal{M}'$).

Lema 2. *Linearna funkcionela L data sa* (1.2) *je slabo neprekidna na $\mathcal{M}$ ako i samo ako može biti reprezentovana kao*

$$L(F) = \int \psi dF \tag{1.2}$$

za neku ograničenu i neprekidnu funkciju ψ.

Definicija 22. Niz $F_n \in \mathcal{M}'$ slabo konvergira ka F ako $\forall \psi \in \mathcal{M}$

$$\langle F_n, \psi \rangle \to \langle F, \psi \rangle.$$

Pišemo $F_n \xrightarrow{w} F$.

Lema 3. *Sledeći iskazi su ekvivalentni:*

1. $F_n \xrightarrow{w} F$.
2. $\liminf F_n\{G\} \geq F\{G\}$ *za svaki otvoren skup G.*
3. $\limsup F_n\{A\} \leq F\{A\}$ *za svaki zatvoren skup A.*
4. $\lim F_n\{B\} = F\{B\}$ *za svaki Borelov skup sa rubom čija je mera F jednaka nuli (odnosno za $F \in \mathcal{M}$ važi $F\{int(B)\} = F\{B\} = F\{\overline{B}\}$).*[8]

Posledica 1. *Ako je F definisana na $\mathbb{R}$ tada slaba konvergencija $F_n \xrightarrow{w} F$ važi ako i samo ako niz funkcija raspodeli F_n konvergira ka F u svakoj tački u kojoj je F neprekidna.*

Napomena: Ova posledica govori o ekvivalenciji slabe konvergencije $F_n \xrightarrow{w} F$ i konvergencije u raspodeli $X_n \xrightarrow{r} X$.

Definicija 23. Podskup $\mathcal{S} \in \mathcal{M}$ je tesan ako $\forall \epsilon > 0$, postoji kompaktan skup $K \subset \Omega$, tako da za svako $F \in \mathcal{S}$ važi $F\{K\} \geq 1 - \varepsilon$.

Lema 4. *Podskup $\mathcal{S} \subset \mathcal{M}$ je tesan ako i samo ako $\forall \epsilon > 0$, $\forall \delta > 0$ postoji konačna unija*

$$B = \bigcup_i B_i$$

zatvorenih δ-lopti, $B_i = \{y \mid d(x_i, y) \leq \delta\}$, tako da za svako $F \in \mathcal{S}$, $F\{B\} \geq 1 - \varepsilon$.

Teorema 8 (Prohorov)**.** *Skup $S \subset \mathcal{M}$ je tesan ako i samo ako je njegovo slabo zatvaranje slabo kompaktno.*

[8] int(B) je unutrašnjost skupa, a $\overline{B}$ zatvaranje skupa B.

1.2.1 Lévi i Prohorov metrika

Sada je potrebno pokazati da je prostor verovatnoća $\mathcal{M}$, indukovan slabom topologijom, definisan na poljskom prostoru Ω i sam poljski prostor.

Definicija 24. Lévi rastojanje između dve funkcije raspodele F i G dato je sa

$$d_L(F,G) = \inf\{\varepsilon > 0 \mid \forall x \in \mathbb{R}\ F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\}. \quad (1.3)$$

U slučaju kada je $\Omega = \mathbb{R}$ ova metrika se najvčešće koristi za metrizaciju prostora mera verovatnoće $\mathcal{M}$.

Lema 5. *Lévi rastojanje dato sa d_L je metrika.*

Dokaz.

1. $d_L(F,G) \geq 0,\ d_L(F,G) = 0$ ako i samo ako $F = G$.
2. $d_L(F,G) = d_L(G,F)$.
3. $d_L(F,H) \leq d_L(F,G) + d_L(G,H)$.

□

Teorema 9. *Lévi rastojanje metrizuje slabou topologiju.*

Za poljska polja Ω, slaba topologija na $\mathcal{M}$ može biti metrizovana i Prohorov rastojanjem. Prvo će biti definisano par pojmova potrebnih za definiciju ove metrike.

Za svaki podskup $A \subset \Omega$, *zatvorena δ-okolina* skupa A je

$$A^\delta = \{x \in \Omega \mid \inf_{y \in A} d(x,y) \geq \delta\}.$$

Može se pokazati da je A^δ zatvoren skup. Za fiksiranu proizvoljnu raspodelu $G \in \mathcal{M}$ *Prohorov okolina* raspodele G data je sa

$$\mathcal{P}_{\varepsilon,\delta} = \{F \in \mathcal{M} \mid F\{A\} \leq G\{A^\delta\} + \varepsilon\ \ \forall A \in \mathcal{B}\}. \quad (1.4)$$

Definicija 25. Prohorov rastojanje između $F, G \in \mathcal{M}$ dato je sa

$$d_P(F,G) = \inf\{\varepsilon > 0 \mid F\{A\} \leq G\{A^\varepsilon\} + \varepsilon \text{ za svako } A \in \mathcal{B}\}.$$

Lema 6. *Prohorov rastojanje d_P je metrika.*

Teorema 10 (Strassen)**.** *Sledeća dva tvrđenja su ekvivalentna:*

(1) $F\{A\} \leq G\{A^\delta\} + \varepsilon$ *za svako* $A \in \mathcal{B}$.

(2) *Postoje (zavisne) slučajne promenljive X i Y sa vrednostima u Ω tako da $\mathcal{L}(X) = F$ i $\mathcal{L}(Y) = G$ i važi $P\{d(X,Y) \leq \delta\} \geq 1 - \varepsilon$.*

$\mathcal{L}$ je prostor pozitivnih linearnih funkcija na C takvih da je $\mathcal{L}(1) = 1$, gde je C je prostor ograničenih i neprekidnih funkcija na Ω[9].

[9]Poznato je da postoji jednoznačna veza između pozitivnih linearnih funkcionala sa osobinom $\mathcal{L}(1) = 1$ nad $\mathcal{C}$ i izmedju verovatnosnih mera F.

Napomena 1: Pretpostavimo da je dato $\delta = 0$ i neka su raspodele su F i G regularne. Tada (1) iz navedene teoreme implicira da totalna varijacija između F i G zadovoljava

$$d_{TV}(F,G) = \sup_{A\in\mathcal{B}} \big|F\{A\} - G\{A\}\big| \leq \varepsilon.$$

U ovom slučaju Štrasenova teorema govori da postoje dve slučajne promenljive X i Y sa navedenim marginalnim raspodelama tako da $P\{X \neq Y\} \leq \varepsilon$. Odnosno rastojanje d_{TV} ne metrizuje slabu topologiji $\mathcal{M}$.

Napomena 2: Štrasenova teorema pokazuje da uvek možemo pretpostaviti da postoji prava raspodela, koju je nemoguće registrovati, slučajne promenljive Y sa funkcijom raspodele $\mathcal{L}(Y) = G$ i empirijska slučajna promenljiva X, koju je moguće registrovati, sa funkcijom raspodele $\mathcal{L}(X) = F$, tako da $P\{d(X,Y) \leq \varepsilon\} \geq 1 - \varepsilon$. Ova teorema sa Prohorov rastojanjem intuitivno govori i kvantifikuje informacije o malim greškama koje se događaju vrlo verovatno, ali i informacije o velikim greškama koje se događaju sa malim verovatnoćama.

U slučaju kada $\Omega = \mathbb{R}$ često se koristi Kolmogorov rastojanje dato sa

$$d_K(F,G) = \sup_{x\in\mathbb{R}} \big|F(x) - G(x)\big|.$$

Ni rastojanje totalne varijanse d_{TV}, ni Kolmogorov rastojanje d_K ne generišu slabu topologiju na $\mathcal{M}$.

Može se pokazati

$$2d_{TV}(F,G) = \int_{\chi} |dF - dQ|.$$

Primer 3. *Neka je data eksponencijalna raspodela $F : \varepsilon(1)$ i neka je data uniformna raspodela G na $\mathbb{R}(0,1)$. Onda*

$$2d_{TV}(F,G) = \int_0^1 (1 - e^{-x})dx + \int_1^\infty e^{-x}dx = \frac{1}{e} = \frac{2}{e}$$

Dalje, Kolmogorov rastojanje dato je sa

$$d_K(F,G) = \sup_{x\geq 0} \big|1 - e^{-x} - xI[0 \leq x \leq 1] - I[x > 1]\big| = e^{-1}.$$

Primer 4. *Neka je data binomna raspodela $F : \mathcal{B}(n = 100, p = 0.001)$ i neka je data poasonova raspodela $G : P(\lambda = 1)$. Onda je*

$$2d_{TV}(F,G) = \sum_{k=0}^{100} \left| \binom{n}{k} 0.01^k 0.99^{100-k} - \frac{e^{-1}}{k!} \right| + \sum_{k=101}^{\infty} \frac{e^{-1}}{k!}.$$

Ovo je relativno lako izračunati u R-u komandom

> sum(abs(dbinom(0:100,100,0.01)-dpois(0:100,1))) +1-ppois(100,1)

[1] 0.005550589.

Odavde sledi da je $d_{TV}(F,G) \approx 0.0028$. *Dalje, Kolmogorov rastojanje dato je sa*

$$d_{TV}(F,G) = \max_{k\in N_0}\left|\sum_{k=0}^{100}\binom{n}{k}0.01^k 0.99^{100-k} - \sum_{k=0}^{\infty}\frac{e^{-1}}{j!}\right|,$$

odnosno

```
> max(abs(pbinom(0:100,100,0.01)-ppois(0:100,1)))

[1] 0.0018471.
```

Teorema 11. *Prohorov metrika* d_P *metrizuje slabu topologiji definisanu na* $\mathcal{M}$.

Teorema 12. $\mathcal{M}$ *je poljski prostor.*

Ograničena Lipšic metrika

Slaba topologija može biti metrizovana i drugim metrikama. Pretpostavimo da je razdaljina d na Ω ograničena sa 1. Ovo sledi iz teoreme.

Teorema 13. *Neka je* (X,d) *metrički prostor. Tada važi:*

1. *Funkcija* $d_l : X^2 \mapsto \mathbb{R}$ *data sa* $d_l(x,y) = \frac{d(x,y)}{1+d(x,y)}$ *je ograničena metrika na skupu* X, *jer je* $d_l(x,y) < 1$.
2. *Metrike* d *i* d_l *na skupu* X *određuju istu topologiju.*

Definicija 26. Ograničena Lipšic metrika data je sa

$$d_{BL}(F,G) = \sup\left|\int \psi dF - \int \psi dG\right|,$$

gde je supremum uzet nad svim funkcijama koje zadovoljavaju Lipšicov uslov, odnosno

$$\left|\psi(x) - \psi(y)\right| \leq d(x,y).$$

Lema 7. d_{BL} *je metrika.*

Teorema 14. *Sledeća dva iskaza su ekvivalentna:*

1. $d_{BL}(F,G) \leq \varepsilon$.
2. *Postoje dve slučajne promenljive* X *i* Y *sa funkcijama raspodele* $\mathcal{L}(X) = F$ *i* $\mathcal{L}(Y) = G$ *tako da*
$$E(d(X,Y)) \leq \varepsilon.$$

Posledica 2. *Za svako* $F,\ G \in \mathcal{M}$ *važi*

$$d_P(F,G)^2 \leq d_{BL}(F,G) \leq 2d_P(F,G).$$

Odnosno, d_P *i* d_{BL} *metrizuju istu topologiju.*

1.2.2 Fréchet i Gâteaux izvod

Neka je d_* proizvoljna metrika na prostoru $\mathcal{M}$ tako da:

(1) Je kompatibilna sa slabom topologijom definisanom na $\mathcal{M}$ u smislu da je skup $\{F \mid d_*(G,F) < \varepsilon\}$ je otvoren za svako $\varepsilon > 0$.

(2) Je kompatibilna sa afinom strukturom $\mathcal{M}$. Ako $F_t = (1-t)F_0 + tF_1$, onda važi $d_*(F_t, F_s) = O(|t-s|)$[10].

Navedene metrike očigledno zadovoljavaju osobinu (1), dok je osobinu (2) potrebno ispitati. U slučaju Prohorov metrike dobija se

$$\left|F_t\{A\} - F_s\{A\}\right| = |t-s|\left|F_1\{A\} - F_0\{A\}\right| \leq |t-s|\,.$$

Osobina (2) trivijalno sledi i za ostale navedene metrike.

Definicija 27. Funkcionela T je Fréchet diferencijabilna u F ako može biti aproksimiriana linearnom funkcijom L_F, definisanom na prostoru svih konačnih mera sa predznakom, tako da za svako G

$$\left|T(G) - T(F) - L_F(G-F)\right| = o(d_*(F,G)).^{11} \tag{1.5}$$

Može se pokazati da je linearna funkcija L_F iz prethodne definicije jedinstvena na prostoru svih konačnih mera totalne algebarske mase 0[12]. Takođe ona može biti standardizovana tako da $L(F) = 0$.

Ako je T definisana na otvorenoj okolini F na nekom linearnom prostoru, onda slaba neprekidnost T u F zajedno sa (1.5) implicira da je L_F neprekidna po G u $G = F$ i budući da je L_F linearna sledilo bi neprekidnost L_F svuda. Nažalost funkcionela T se uglavnom definiše na nekim konveksnim skupovima.

Definicija 28. Nosač $\operatorname{supp}\varphi$ funkcije $\varphi : \mathbb{R} \mapsto C$, je zatvaranje skupa svih tačaka $t \in \mathbb{R}$ u kojima je $\varphi(t) \neq 0$, tj.

$$\operatorname{supp}\varphi = \overline{\{t \in \mathbb{R} : \varphi(t) \neq 0\}}.$$

Definišimo $\psi(x) = L(\delta_x - F)$, onda zbog linearnosti L (i mogućnosti standardizacije) sledi

$$L(G-F) = \int \psi dG$$

za sve G sa konačnim nosačem. Ako je

$$F_t = (1-t)F + tG, \tag{1.6}$$

onda,

$$\begin{aligned} \left|T(F_t) - T(F) - L(F_t - F)\right| &= \left|T(F_t) - T(F) - tL(G-F)\right| \\ &= \left|T(F_t) - T(F) - t\int \psi dG\right| = o(d_*(F,F_t)) = o(t). \end{aligned} \tag{1.7}$$

[10] $f(x) = \mathcal{O}(g(x)) \Leftrightarrow f(x) \leq Mg(x)$

[11] $f(x) = o(g(x)) \Leftrightarrow \lim \frac{f(x)}{g(x)} = 0$

[12] Mere koje svakom skupu dodeljuju meru 0.

Pretpostavimo da je T neprekidna u F, onda $d_*(F, F_t) = O(t)$ što implicira da $\left|T(F_t) - T(F)\right| = o(1)$ konvergira unfirmno na G. Takođe iz (1.7) vidimo da ψ mora biti ograničena.

Posledica 3. *Ako je T slabo neprekidna u okolini F i ako je Fréchet diferencijabilna u F, onda je njen Fréchetov (Frešov) izvod u F slabo neprekidana linearna funkcija L_F i može se zapisati kao*

$$L_F(G - F) = \int \psi_F dG$$

gde je ψ_F ograničena i neprekidna funkcija tako da važi $\int \psi_F dF = 0$.

Nažalost ovaj diferencijal često nije definisan ili ako postoji često je teško ustanoviti njegovo postojanje, zato se često koriste nešto slabiji diferencijali.

Definicija 29. Funkcionela T je Gáteaux diferencijabilna u raspodeli F ako postoji realna i merljiva funkcija ψ_F tako da za svako $G \in \mathcal{M}$ važi da

$$\lim_{t \to 0} \frac{T((1-t)F + tG) - T(F)}{t} = \int \psi_F dG(x),$$

odnosno Gáteaux izvod je upravo izvod funkcionele T u F po pravcu G. Ovo se ekvivalentno može zapisati kao

$$\frac{\partial}{\partial t}[T((1-t)F + tG) - T(F)]_{t=0} = \int \psi_F dG(x).$$

Kroz rad se pretpostavlja da je domen T konveksan skup i generalno se bira funkcija ψ_F tako da važi $\int \psi_F dF(x) = 0$. Gáteaux diferencijal je upravo običan diferencijal realne funkcije $T(F_t)$ gde je F_t dato sa (1.6) i $G = \delta_x$ u odnosu na promenljivu t.

1.2.3 Hampelova teorema

Neka su observacije x_i *i.i.d.* sa funkcijom raspodele F i neka je $T_n = T(x_1, ..., x_n)$ niz ocenjivača sa vrednostima iz $\mathbb{R}$.

Definicija 30. Niz ocenjivača $\{T_n\}_{n \in N} \in R^k$ je (kvantitativno) robusan u F ako je niz preslikavanja $F \mapsto \mathcal{L}_F(T_n)$ podjednako neprekidan.

Definicija 31. Preslikavanje je podjednako neprekidno ako $\forall \varepsilon > 0,\ \exists \delta > 0$ i $\exists n_0$ tako da za $\forall G \in \mathcal{M}$ i $\forall n \geq n_0$ važi

$$d_*(F, G) \leq \delta \to d_*(\mathcal{L}_F(T_n), \mathcal{L}_G(T_n)) \leq \varepsilon.$$

Gde je d_* metrika koja metrizuje slabu topologiju. Nadalje se koristi Lévijeva metrika za F i Prohorova metrika za $\mathcal{L}(T_n)$. Nije poznato da li različite metrike daju ekvivalentne pojmove robusnosti.

Pretpostavimo da je $T_n = T(F_n)$ potiče od funkcionele T koja je definisana na nekom slabo otvorenom skupu iz $\mathcal{M}$.

Posledica 4. *Ako je T slabo neprekidna u F, onda $\{T_n\}_{n\in N}$ je konzistentna u F u smislu da $T_n \xrightarrow{p} T(F)$ i $T_n \xrightarrow{s.s.} T(F)$.*

Teorema 15 (Hampel)**.** *Pretpostavimo da $\{T_n\}$ potiče od funkcionele T i da je konzistentan u okolini funkciji raspodele F. Onda je T neprekidna u F ako i samo ako je $\{T_n\}$ robusna u funkciji raspodele F.*

1.3 Tačka preloma statistike

Grubo rečeno tačka preloma je namjanji deo uzorka koji izaziva inflaciju vrednosti ocenjivača. Malo preciznije rečeno to je limit distance od raspodele modela posle koje statistika postaje nepouzdana i neinformativna.

Definicija 32. Tačka preloma ε^* niza ocenjivača $\{T_n\}_{n\geq 1}$ u F je definisana kao

$$\varepsilon^* = \sup\big\{\varepsilon \leq 1 : \text{postoji kompaktan skup } \ K_\varepsilon \subsetneq \Theta$$
$$\text{tako da } \ d_P(F,G) < \varepsilon \Rightarrow G(\{T_n \in K_\varepsilon\}) \xrightarrow{n\to\infty} 1\big\}.$$

Prohorov metrika se može zameniti i drugim (pomenutim) metrikama što daje varijacije ove definicije.

Primer 5. *Neka je $\Theta = \mathbb{R}$, onda sledi da je*

$$\varepsilon^* = \sup\big\{\varepsilon \leq 1 : \textit{postoji kompaktan skup } \ r_\varepsilon \textit{ tako da}$$
$$d_P(F,G) < \varepsilon \Rightarrow G(\{|T_n| \leq r_\varepsilon\}) \xrightarrow{n\to\infty} 1\big\}.$$

Ovaj koncept je kasnije našao primenu na uzorcima konačne populacije gde se ispostavilo da je koristan i igra vrlo važnu ulogu. U literaturi se može pronaći više definicija tačke preloma koje su date u zavisnosti od obima populacije.

Definicija 33. Tačka preloma konačnih uzoraka ε_n^* ocenjivača T_n na uzorku obima n data je sa

$$\varepsilon_n^* = \frac{1}{n}\max\{m : \max_{i_1,\ldots,i_m} \sup_{y_1,\ldots,y_m} \big|T_n(z_1,\ldots,z_n)\big| < \infty\}$$

gde je $(z_1,\ldots,z_n)$ dobijen zamenom m tačaka uzorka $x_{i_1},\ldots,x_{i_m}$ sa proizvoljnim vrednostima $y_1,\ldots,y_m$.

Napominjemo da ovako definisana tačka preloma ne zavisi od uzorka $(x_1,\ldots,x_n)$, već zavisi od obima uzorka n. Takođe napominjemo da je ova definicija prilagođena za ocenjivače parametra lokacije, dok je za parametar skaliranja potrebna mala modifikacija ove definicije, odnosno potrebno je da važi i

$$\min_{i_1,\ldots,i_m} \inf_{y_1,\ldots,y_m} T_n(z_1,\ldots,z_n) > 0.$$

Većina ocenjivača jednog parametra dostiže visoke tačke preloma, iako im ovo nije primarna svrha. Ocenjivači više parametara često nemaju visoku tačku preloma

i često se pojavljuje problematika kod njenog određivanja, upravo zbog više parametara što će biti prezentovano na istovremenim ocenjivačima regresije i skaliranja.

U literaturi se mogu naći metode koje maksimiziraju tačku preloma stavljajući je u centar pažnje modeliranja zanemarujući efikasnost. Ove metode pored toga što su neefikasne, često nailaze na probleme sa stabilnošću, pa se teško mogu zvati robusne. Obrađenim metodama su navedene odgvoarajuće prelomne tačke bez preteranih analiza.

1.4 Kvalitativna robusnost

Neka je F_n empirijska funkcija raspodele slučajne promenljive X i definisana je kao

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \leq x\}},$$

gde je indikator funkcija definisana kao $I = \begin{cases} 1, & x_i \leq x, \\ 0, & x_i > x. \end{cases}$

Definicija 34. Ako $F_n \xrightarrow{p} F$, kažemo da je funkcionela T Fišer konzistentna u F ako važi

$$\lim_{n\to\infty} T(F_n) \xrightarrow{p} T(F).$$

Već je pomenuto u Poglavlju 1.2.3 da je niz ocenjivača $\{T_n\}_{n\in N} \in R^k$ je kvantitativno robusan u F ako je niz preslikavanja $F \to \mathcal{L}_F(T_n)$ podjednako neprekidan.

Definicija 35. Niz ocenjivača $\{T_n\}_{n\in N}$ je neprekidan u F ako za $\forall \varepsilon > 0,\ \exists \delta > 0$ i $\exists n_0$ tako da $\forall n, m \geq n_0$ i za svaku empirijsku funkciju raspodele F_n, F_m važi

$$\left.\begin{array}{l} d_P(F, F_n) < \delta \\ d_P(F, F_m) < \delta \end{array}\right\} \quad \Longrightarrow \quad \left|T_n(F_n) - T_m(F_m)\right| < \varepsilon.$$

1.5 Kvantitivna robusnost

U ovom poglavlju biće reč o kvantifikaciji uticaja koju promena u raspodeli ocenjivača $T(F)$ ima na promenu raspodela ocenjivača, odnosno promenu $\mathcal{L}_F(T_n)$.

Dve okoline na kojim najčešće radimo su Lévi okruženje i ε-kontaminirana okolina raspodele.

Definicija 36. Lévi okruženje se definiše kao

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid \forall x \in \mathbb{R},\ F_0(x-\varepsilon) - \varepsilon \geq F(x) \geq F_0(x+\varepsilon) + \varepsilon\}, \tag{1.8}$$

dok se model ε-kontaminirane okoline (raspodele) definiše kao

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid F = (1-\varepsilon)F + \varepsilon H,\ H \in \mathcal{M}\}, \tag{1.9}$$

gde je H proizvodna raspodela iz neke klase.

Definicija 37. Maksimalna pristrasnost definisana je kao

$$b_{max}(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} \left| T(F) - T(F_0) \right|. \tag{1.10}$$

Definicija 38. Maksimalna varijansa definisana je kao

$$v_{max}(\varepsilon) = \sup_{F \in \mathcal{P}_\varepsilon} V(F, T). \tag{1.11}$$

Često se posmatra supremum $V(F,T)$ na preseku skupa $\mathcal{P}_\varepsilon$ na kome je $T(F)$ konstantno. Priloženim pojmovima se za dovoljno veliko n može pokazati da se ocena T_n ponaša očekivano za fiksno $F \in \mathcal{P}_\varepsilon$. Pristup je nešto drugačiji kada se ispituje ponašanje T_n za svako F iz okoline $\mathcal{P}_\varepsilon$.

Neka je $M_e(F, T_n)$ medijana raspodele $\mathcal{L}_F[T_n - T(F_0)]$ i neka je $Q_t(F, T_n)$ normalizovani t-kvantil opseg (interval koji sadrži $(1-2t)\%$ podataka) raspodela $\mathcal{L}_F(\sqrt{n}T_n)$. Normalizovani t-kvantil opseg proizvoljne raspodela G je definisan kao

$$Q_t = \frac{G^{-1}(1-t) - G^{-1}(t)}{\Phi^{-1}(1-t) - \Phi^{-1}(t)},$$

gde je $\Phi : \mathcal{N}(0,1)$ raspodela, $\Phi(q_1) = t$ i $\Phi(q_2) = 1 - t$.

Vrednost t je proizvoljna, ali fiksirana. Često se koristi $t = 0.25$ (interkvartilni opseg) ili $t = 0.025$ (95% opseg) koji je pogodan za konstrukciju 95% intervala poverenja.

Definicija 39. Maksimalna asimptotska pristrasnost definisana je kao

$$b(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} \left| M_e(F, T_n) \right|. \tag{1.12}$$

Definicija 40. Maksimalna asimptotska varijansa definisana je kao

$$v(\varepsilon) = \lim_n \sup_{F \in \mathcal{P}_\varepsilon} Q_t(F, T_n)^2. \tag{1.13}$$

Teorema 16. *Ako su b_{max} i v_{max} dobro definisani, onda važi $b(\varepsilon) \geq b_{max}(\varepsilon)$ i $v(\varepsilon) \geq b_{max}(\varepsilon)$.*

Iz praktičnih razloga uglavnom se radi sa izrazima (1.10) i (1.11) tj. maksimalnom pristrasnosti i maksimalnom varijansom, s' tim što se za konkreno $\mathcal{P}_\varepsilon$ i T mora proveravati da li važi $b_{max}(\varepsilon) = b(\varepsilon)$ i $v_{max}(\varepsilon) = v(\varepsilon)$, što često važi.

Definicija 41. Asimptotska tačka preloma statistike T za fiksnu raspodelu F_0 definiše se kao

$$\varepsilon^* = \varepsilon^*(F_0, T) = \sup\{\varepsilon \mid b(\varepsilon) < b(1)\}.$$

Grubo govoreći ovako definisana tačka preloma stavlja granicu na količinu loših observacija sa kojima ocenjivač T može da izađe na kraj.

Osetljivosti statistike na jednu observaciju može se kvantifikovati na više načina. Navodimo tri načina gde se prvi odnosi na asimptotski slučaj (obim uzorka $n \to \infty$), a druga dva slučaja na uzorke konačnog obima.

Tabela 1.2: Razne osobine ocenjivača

Funkcionela	$\rightarrow$	Uticajna funkcija
Asimptotska varijansa	$\rightarrow$	Funkcija promene varijanse (CVF)
Ukupna osetljivost na greške	$\rightarrow$	Funkcija promene pristrasnosti (CBF)

Definicija 42. Uticajna funkcija[13] kvantifikuje uticaj pojedinačne observacije x na statistike $T(F)$ u F i definisana je kao

$$IF(x, F, T) = \lim_{s \to 0} \frac{T((1-s)F + s\delta_x) - T(F)}{s}. \tag{1.14}$$

Ukoliko je uticajna kriva neograničena onda je za očekivati da jedan autlajer može učiniti statistiku u potpunosti nepouzdanom. Uticajna kriva nam omogućava da definišemo jos moćniju aparaturu koja opisuje robusnost statistika, što se i može videti na priloženom grafiku.

Za svaku funkcionelu definisana je uticajna funkcija koja dalje definiše asimptotsku varijansu i ukupnu osetljivost na greške. Teorija se znatno komplikuje i stoga su u ovom radu obrađena samo zelena polja koja su potrebna za definisanje osobine *B-robusnosti* i posmatraćemo isključivo *B-robusne* statistike. Drugi pristup je preko analize varijanse i on definiše *V-robusnost*. Analogno analizi varijanse preko uticajne funkcije sada analizira se varijansa ocenjivača.

Definicija 43. Funkcija promene varijanse (*CVF*)[14] data je sa

$$CVF(x, T, F) = \lim_{\varepsilon \to 0} \frac{V(T, (1-\varepsilon)F + \varepsilon\delta_x) - V(T, F)}{\varepsilon}.$$

Analogno se definiše i funkcija promene pristrasnosti. Napominjemo samo da postoje i mere osetljivosti funkcija promene varijanse i pristrasnosti koje govore o stabilnosti ocenjivača. Detaljnije o ovome kao i vezama koje postoje između *B-robusnosti* i *V-robusnosti* u [13]. Nas prvenstveno interesuju kvalitativno robusni ocenjivači koji poseduju visoku tačku preloma i nisku ukupnu osetljivost na greške.

Lema 8 (Glivenko–Cantelli)**.** *Neka su $X_1, ..., X_n$ i.i.d. slučajne promenljive sa funkcijom raspodele $F_X(x)$ i neka je $F_n(x)$ empirijska raspodela funkcije. Onda kada $n \to \infty$ važi*

$$P\big[\sup_{x \in R} \big|F_n(x) - F_X(x)\big| \to 0\big] = 1,$$

ili ekvivalentno

$$P\big[\lim_{n \to \infty} \sup_{x \in R} \big|F_n(x) - F_X(x)\big| = 0\big] = 1,$$

odnosno konvergencija je uniformna po x.

Ako se neka raspodela G *nalazi* u okolini F onda Tejlorov razvoj daje linearizaciju $T(G)$ u okolini $T(F)$ tj.

$$T(G) - T(F) = \int IF(x, F, T) d(G - F)(x) + R,$$

[13] Influence curve ili Influence function.
[14] change-of-variance function.

gde je R neki ostatak. Neka je T Gáteaux diferencijabilna u raspodeli F na domenu funkcionele T, onda važi $\int \psi_F dF = 0$.

Ako $n \to \infty$, pa važi Glivenko-Kanteli lema i G ćemo aproksimirati sa empirijskom funkcijom raspodele F_n tj. $G = F_n$ i ako je T dovoljno regularna, onda sledi

$$T(F_n) - T(F) = \int IF(x, F, T) dF_n(x) + R,$$

odnosno, ako integral sa desne strane ocenimo dobija se

$$\sqrt{n}(T(F_n) - T(F)) \simeq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(x_i, F, T) + R_n.$$

Pod pretpostavkom da su x_i *i.i.d.* centralna granična teorema garantuje da je desna strana ove jednakosti asimptotski normalna sa očekivanjem 0, dok ostatak postaje zanemarljiv kada $n \to \infty$. Može se zaključiti da je T_n asimptotski normalna, odnosno da $\mathcal{L}_F(\sqrt{n}[T(F_n) - T(F)])$ teži ka $\mathcal{N}(0, V(T, F))$ gde je varijansa jednaka

$$V(F, T) = \int IC(x, F, T)^2 dF(x). \tag{1.15}$$

O uslovima regularnosti pod kojim ovo važi više se može naći u [4] strana 40 i [8] strana 85.

Definicija 44. Hampel definiše ukupnu osetljivost[15] na greške kao

$$\gamma^*(F, T) = \sup_{x \in \mathbb{R}} \left| IF(x, F, T) \right|.$$

Ukoliko je $\gamma^*(F, T)$ konačna, odnosno ocenjivač T ima ograničenu uticajnu funkciju, kažemo da je ocenjivač *B-robusan*[16].

Za model kontaminiranih ε-okolina 1.9 važi sledeće

$$T(F) - T(F_0) \approx \varepsilon \int IC(x, F, T) dH(x),$$

odakle sledi

$$b_{max}(\varepsilon) = \sup_x \left| T(F) - T(F_0) \right| \cong \varepsilon \gamma^*.$$

Ukupna osetljivost na greške je centralna lokalna mera robusnosti koja meri maksimalnu pristrasnost izazvanu malim kontaminacijama. Samim tim, može se posmatrati stabilnost statistike T pod malim promenama u raspodeli F. Takođe ovu veličinu možemo posmatrati kao gornju granicu na (standardizovanoj) asimptotskoj pristrasnosti ocenjivača. Poželjno je da ova veličina bude konačna.

Često se prvi korak ka robusifikaciji određenog ocenjivača upravo vrši ograničavanjem ukupne osetljivost na greške, $\gamma^*(F, T)$, što često kao posledicu ima pogoršanje asimptotske efikasnosti ocenjivača.

[15] Gross error sensitivity.
[16] Skraćeno od bias.

Definicija 45. Matrica Fišerove informacije slučajne promenljive X date funkcijom gustine $f(x)$ definisana je kao

$$I(F_\theta) = E\Big[\Big(\frac{\partial}{\partial\theta} log f(X;\theta)\Big)^2 |\theta\Big] = \int \Big(\frac{\partial}{\partial\theta} log f(x;\theta)\Big)^2 f(x;\theta)dx.$$

Definicija 46. Apsolutna asimptotska efikasnost ocenjivača data je sa

$$e := [V(T,F)I(F)]^{-1}$$

gde je $I(F)$ Fišerova matrica informacija.

Definicija 47. Lokalna promena pomeraja definisana je kao

$$\lambda^* = \sup_{x\neq y} \frac{\big|IF(y,T,F) - IF(x,T,F)\big|}{|y-x|}.$$

To je najmanja Lipšic konstanta koju IF zadovoljava.

Ovaj pokazatelj daje informacije o dešavanjima koja nastaju pomeranjem tačke x na njenu okolinu y. Može biti korisna kada se ispituju grešeke na uzorku nastale zaokruživanjem.

Definicija 48. Ukoliko je raspodela F simetrična onda se može definisati tačka odbacivanja kao

$$\rho^* = \inf\{r > 0 \ : \ IF(x,T,F) = 0, \ \forall |x| > r\}. \tag{1.16}$$

Sve observacije dalje od ρ^* se kompletno odbacuju. Poželjno je da ovaj pokazatelj bude konačan. Posmatrajmo regiju tačaka za koje je $IF(x,F,T) = 0$. To znači da ove tačke nemaju uticaja na ocenjivač pa ni njihova kontaminacija ne utiče na ocenjivač.

U slučaju konačnog uzorka navedene su kriva osetljivosi i *jackknife*.

Definicija 49. Standardizovana kriva osetljivosti (Taki[17], 1970) je definisana kao

$$SC_{n-1}(x) = \frac{T(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_x) - T(F_{n-1})}{\frac{1}{n}}$$
$$= n[T_n(x_1,...,x_{n-1},x) - T_{n-1}(x_1,...,x_{n-1})],$$

i postoje dve verzije ove krive, jedna sa dodavanjem nove observacije x a druga sa zamenom postojeće observacije x_i observacijom x.

Croux (1998) pokazuje da ako je T_n *M* ocenjivač za lokaciju sa ograničenom i neprekidnom funkcijom ψ ili odsečeno očekivanje[18] onda za svako x važi

$$SC_{n-1}(x) \xrightarrow{s.s.} IF(x,F,T_n).$$

Naglašavamo da ovo važi samo u okolini tačke x za dovoljno veliko n koje zavisi od x.

[17] John Tukey.
[18] Trimmed mean.

Definicija 50. *Jackknifed pseudovrednost* i-tog *opažanja* je definisana kao

$$T_{ni}^* = nT_n - (n-1)T_{n-1}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n).$$

Generalno statistika data funkcionelom T na uzorku $(x_1, ..., x_n)$ daju pseudouzorak $(T_{n1}^*, ..., T_{nn}^*)$ koji se koristi za računanje korigovanog ocenjivača

$$T_n^* = \frac{1}{n}\sum_{i=1}^{n} T_{ni}^*$$

koji često ima manju pristrasnost od početnog ocenjivača T_n. Varijansa ovih n pseudovrednosti je

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(T_{ni}^* - T_n^*)^2.$$

Taki navodi da je $\frac{1}{n(n-1)}\sum(T_{ni}^* - T_n^*)^2$ dobra aproksimacija za varijansu T_n.

Ukoliko je je T_n uzoračko očekivanje, onda sledi $T_{ni}^* = x_i$.

Primer 6. *Neka je u izrazu* (1.14) $s = -\frac{1}{n-1}$ *i neka je* $F = F_n$ *može se uočiti sledeće*

$$\begin{aligned} IF(F_n, T, x_i) &= \lim_{n\to\infty} \frac{T(\frac{n}{n-1}F_n - \frac{1}{n-1}\delta_{x_i}) - T(F_n)}{\frac{-1}{n-1}} \\ &= \lim_{n\to\infty}(n-1)[T_n - T_{n-1}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)] \\ &= T_{ni}^* - T_n. \end{aligned}$$

Primer 7. *Neka je* $\mathcal{B} = \mathbb{R}$ *i neka je* $\Theta = \mathbb{R}$. *Posmatra se model lokacije dat sa* $F_\theta(x) = \Phi(x-\theta)$, *gde je* Φ *standardna normalna funkcija raspodele. Radi jednostavnosti pretpostavimo da je* $\theta = 0$, *odakle sledi da je* $F_\theta = \Phi$. *Razmotrimo aritmetičko očekivanje dato statistikom* $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ *i odgovarajućom funkcionelom* $T(G) = \int u dG(u)$ *i ono postoji za svaki prostor verovatnoća kada postoji prvi momenat. Očigledno* T *je Fisher konzistentna. Uticajna funkcija ove statistike je*

$$\begin{aligned} IF(x, T, \Phi) &= \lim_{s\to 0} \frac{\int u\, d[(1-s)\Phi + s\delta_x](u) - \int d\Phi(u)}{s} \\ &= \lim_{s\to 0} \frac{(1-s)\int u d\Phi(u) + s\int u\delta_x(u) - \int u d\Phi(u)}{s} \\ &= \lim_{s\to 0} \frac{sx}{s} = x. \end{aligned}$$

Ovo sledi iz $\int u d\Phi(u) = 0$. *Primetimo* $\int IF(x, T, \Phi)d\Phi(x) = 0$. *Ukupna osetljivost na greške je*

$$\gamma^*(\Phi, T) = \sup_x \big|IF(x, \Phi, T)\big| = \sup_x |x|$$

neograničena što znači da ova statistika nije B-robusna. *Takođe primetimo*

$$\int IF(x, T, F)d\Phi(x) = \int x d\Phi(x) = 0,$$

a varijansa

$$V(T,\Phi)=\int IF(x,T,F)^2 d\Phi(x)=\int x^2 d\Phi(x)=1.$$

Budući da je uzoračko očekivanje ML ocenjivač onda je asimptotska efikasnost modela $e=1$. *Tačka odbacivanja ne postoji, odnosno,* $\rho^*=\infty$. *Tačka preloma uzoračkog očekivanja je* $\varepsilon^*=0$ *i ovaj ocenjivač nije kvalitativno robusan. Grafik uticajne funkcije ove statistike dat je na Slici 1.1. Lokalna promena pomeraja data je sa*

$$\lambda^*=\sup_{x\neq y}\frac{\big|IF(y,T,F)-IF(x,T,F)\big|}{|y-x|}=1.$$

Dakle, možemo zaključiti da ova statistika nije osetljiva na lokalne promene pomeraja.

Primer 8. *Neka važe uslovi iz prethodnog primera. Sada će biti analizirana uzoračka medijana. Neka je obim uzorka* n, *ako je* n *neparan broj onda* $T_n=x_{\frac{n+1}{2}}$, *inače je* $T_n=x_{\frac{n}{2}}+x_{\frac{n}{2}+1}$ *i odgovarajućom funkcionelom* $T(G)=G^{-1}(\frac{1}{2})$ *gde se u slučaju nejedinstvenosti medijane uzima sredina intervala. Ova funkcionela je Fišer konzistentna. Uticajna funkcije ove statistike je*

$$IF(x,T,\Phi)=\frac{\operatorname{sign}(x)}{2\Phi(0)}.$$

Primetimo $\int IF(x,T,\Phi)d\Phi(x)=\int\frac{\operatorname{sign}(x)}{2\Phi(0)}=0$. *Ukupna osetljivost na greške je*

$$\gamma^*(\Phi,T)=\sup_x\big|IF(x,\Phi,T)\big|=\sup_x\left|\frac{\operatorname{sign}(x)}{2\Phi(0)}\right|=\sqrt{\frac{\pi}{2}}$$

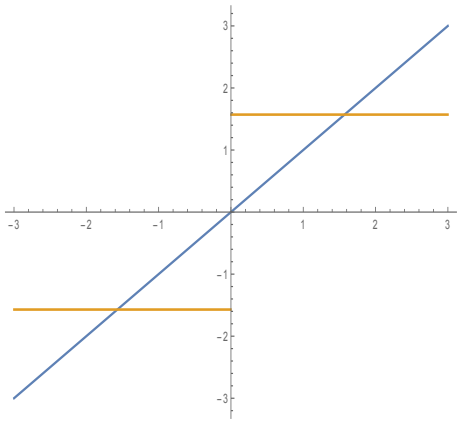
ograničena što znači da je ova statistika B-robusna. *Takođe primetimo da je*

$$V(T,\Phi)=\int(\frac{\operatorname{sign}(x)}{2\Phi(0)})^2 d\Phi(x)=\frac{\pi}{2}=1,571.$$

Asimptotska efikasnost je $e=[V(T,F)I(F)]^{-1}=\frac{2}{\pi}$, *a tačka odbacivanja ne postoji, odnosno* $\rho*=\infty$. *Tačka preloma medijane je* $\varepsilon^*=\frac{1}{2}$ *i ona je kvalitativno robusna ako* $\Phi^{-1}(\frac{1}{2})$ *je jedinstvena tačka. Grafik uticajne funkcije ove statistike dat je na Slici 1.1. Lokalna promena pomeraja data je sa*

$$\begin{aligned}\lambda^*=\sup_{x\neq y}\frac{\big|IF(y,T,F)-IF(x,T,F)\big|}{|y-x|}&=\sup_{x\neq y}\frac{\left|\frac{\operatorname{sign}(y)}{2\Phi(0)}-\frac{\operatorname{sign}(x)}{2\Phi(0)}\right|}{|y-x|}\\&\leq\sup_{x\neq y}\frac{1}{\Phi(0)}\frac{1}{|y-x|}.\end{aligned}$$

Dakle možemo zaključiti da je uzoračka medijana osetljiva na lokalne promene jer $g(x,y)=\frac{1}{|y-x|}$ *nije ograničena funkcija.*



Slika 1.1: Uticajne funkcije statistika uzoračkog očekivanja (plava boja) i uzoračke medijane (žuta)

Posmatrajući ove primere sasvim se prirodno nameće ideja za reopadajuće ocenjivače iz tri dela (o kojima će biti reči kasnije) koji imaju konačnu ukupnu osetljivost na greške γ^*, su mnogo efikasniji od medijane i imaju konačnu tačku odbacivanja ρ^*.

U narednom primeru navodimo još jednu interesantnu robusnu statistike za uzoračko očekivanje (lokaciju) baziranu na ideji odsecanja autlajera koja je iz klase L ocenjivača. Dalje neće biti reči o ovoj klasi ocenjivača.

Primer 9 (Odsečeno očekivanje)**.** *L ocenjivač α-trimovanog ocekivanja, gde je $(0 < \alpha < \frac{1}{2})$, dato je sa*

$$T(F) = \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(t)dt,$$

odnosno u obliku

$$\overline{x_\alpha} = \frac{1}{n-2m}\sum_{i=m+1}^{n-m} x_{(i)},$$

gde je $m = [(n-1)\alpha]$ ceo deo, a $x_{(i)}$ je statistika i-tog reda. Prelomna tačka ove statistike je $\varepsilon^* = \alpha$.

Upravo zbog načina na koji je definisan ovaj ocenjivač i on poseduje ograničenu uticajnu funkciju, odnosno on je *B-robusan*.

Primer 10. *Ocenjivač varijanse dat je sa*

$$Var[X] = S(F) = \int_{\mathbb{R}} x^2 dF(x) - E^2[x],$$

odnosno na konkretnom uzorku

$$S(F_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2.$$

Neka je m_F pravo očekivanje raspodele F, a σ_F^2 prava varijansa ove raspodele. Uticajna funkcija ove statistike data je sa

$$IF(x, S, F) = (x - m_F)^2 - \sigma_F^2.$$

Primetimo $\int IF(x,S,F)dF(x)=0$. *Ukupna osetljivost na greške je*

$$\gamma^*(F,S)=\sup_x\left|IF(x,S,F)\right|=\sup_x\left|(x-m_F)^2-\sigma_F^2\right|$$

neograničena što znači da ova statistika nije B-robusna.

$$V(S,F)=\int IF(x,S,F)^2dF(x)=V_F[(x-m_F)^2-\sigma_F^2]=V_F[x^2-2m_Fx]$$
$$=E_F[x^4]-E_F^2[x^2]+4m_F^2V[x]=\ldots=E[x^4]-\sigma_F^4.$$

Buduci da je uzoračko očekivanje ML ocenjivač onda je asimptotska efikasnost modela $e=1$. *Tacka odbacivanja ne postoji, odnosno,* $\rho^*=1$. *Lokalna promena pomeraja data je sa*

$$\lambda^*=\sup_{x\neq y}\frac{\left|(y-m_F)^2-(x-m_F)^2\right|}{|y-x|}=\ldots=\sup_{x\neq y}|x+y-2m_F|\,,$$

odakle sledi da je ova statistika osetljiva na lokalne promene pomeraja.

Primer 11. *U ovom primeru izvodimo uticajnu funkciju varijanse ocenjivača varijanse. Fukcionela* S *data je sa*

$$S((1-t)F+tG)=\int_{\mathbb{R}}x^2d((1-t)F+tG)-\Big[\int_{\mathbb{R}}xd((1-t)F+tG)\Big]^2$$

Definišemo

$$h(t)=(1-t)E_F[X^2]+tE_G[X^2]-(1-t)^2E_F[X]^2$$
$$-t^2E_G[X]^2-2t(1-t)E_F[X]E_G[X].$$

Sledi,

$$h'(t)=-E_F[X^2]+E_G[X^2]+2(1-t)E_F[X]^2$$
$$-2tE_G[X]^2-2(1-2t)E_F[X]E_G[X].$$

Dalje, uticajna funkcija je izvod funkcionele S *po pravcu* G, *odnosno*

$$\lim_{t\to0}h'(t)=S'_G(F)=-E_F[X^2]+E_G[X^2]+2E_F[X]^2-2E_F[X]E_G[X].$$

Pa za $G=\delta_x$ *sledi*

$$S'_x(F)=x^2-E_F[X^2]-2xE_F[X]+2E_F[X]^2.$$

Odnosno,

$$S'_x(F)=(x-E_F[X])^2-Var_F[X].$$

Glava 2

Klasa *M* ocenjivača

Klasa *M* ocenjivača je *robusna* generalizacija maksimalno verodostojnih ocenjivača i uvodi je Peter J. Huber 1964 godine. Postoje i druge robusne klase ocenjivača poput L, R i S ocenjivača. Klasa ocenjivača tipa maksimalne verodostojnosti bira onu vrednost nepoznatog parametra $\theta \in \Theta$ koja daje najveću verovatnoću da baš taj uzorak bude odabran. Specijalno, u zavisnosti od izbora pretpostavki modela u ovoj klasi se nalaze i neki dobro poznati ocenjivači, što će biti pokazano kasnije. Klasa *M* ocenjivača je prilično fleksibilna i minimizira maksimalnu degradaciju tj. za najmanje povoljnu situaciju predlaže najbolje rešenje.

U ovoj glavi uvodimo *M* ocenjivače parametra lokacije i skale. Na ovim jednostavnim modelima ilustrujemo osnovnu teoriju i ideje iza robusnih ocenjivača. Regresija će biti posebno obrađena u Glavi 3.

Definicija 51. Ocenjivač je skalarno translatoran ako važi

$$T_n = (x_1 + c, ..., x_n + c) = c + T_n(x_1, ..., x_n).$$

Definicija 52. Ocenjivač je skalarno invarijantan ako važi

$$T_n = (cx_1, ..., cx_n) = c\, T_n(x_1, ..., x_n).$$

Primer 12. *Uzoračko očekivanje, odsečeno uzoračko očekivanje i medijana su skalarno translatorni i skalarno invarijantni.*

Definicija 53. ρ - tip *M* ocenjivača za parametar $\theta \in \Theta$ je statistika $T_n = T_n(X_1, ..., X_n)$, odnosno $T_n = T(F_n)$ koja je rešenje problema

$$T_n := \arg\min_{T_n \in \Theta} \sum_{i=1}^{n} \rho(X_i, T_n), \tag{2.1}$$

ili u obliku funkcionele kao

$$T(F) := \arg\min_{\theta \in \Theta} \int \rho(x, \theta) dF(x), \tag{2.2}$$

gde je funkcija ρ definisana na $\rho : \Omega \times \Theta \to R$.

Definicija 54. Ukoliko postoji $\psi(x,t) = \frac{\partial}{\partial\theta}\rho(x,\theta)\ \forall x,\theta$, onda ψ - tip *M* ocenjivača za parametar $\theta \in \Theta$ je statistika $T_n = T_n(X_1,...,X_n)$, odnosno $T_n = T(F_n)$ koja je rešenje problema

$$\sum_{i=1}^{n} \psi(X_i, T_n) = 0, \tag{2.3}$$

ili u obliku funkcionele kao

$$\int \psi(x, T(F))dF(x) = 0. \tag{2.4}$$

U zavisnosti od izbora funkcija ρ tj. ψ dobijaju se različiti ocenjivači za parametara modela. Generalno problemi (2.1) i (2.3) nisu uvek ekvivalentni i mogu dati različita rešenja, ali problem (2.3) je neretko jednostavniji za rešavanje i često se može iskoristiti za početno rešavanje (2.1).

Postavlja se pitanje pod kojim uslovima postoji konvergentan niz rešenja problema (2.1) i (2.3). Egzistencija jedinstvenog rešenja (2.1) i (2.3) nije garantovana Glivenko–Cantelli lemom bez dodavanja određenih uslova na funkciju ρ, ψ i/ili funkciju raspodele F. Odgovor na ovo pitanje dat je u narednom delu teksta gde su nešto detaljnije navedene analiza po ρ i ψ kroz par teorema. Više o ovome može se pročitati u literaturi: Huber (1981, Poglavlje 6), Serfling (1980, Poglavlje 7), Lehmann (1983, Poglavlje 6).

Ako je ρ svuda diferencijabilna funkcija i $\psi = \frac{\partial}{\partial\theta}\rho$ monotona funkcija onda su problemi (2.1) i (2.3) ekvivalentni, odnosno daju ista rešenja. Takođe navodimo da sistem (2.3) može da ima više rešenja od kojih je samo jedno rešenje globalni minimum (2.1).

Sistem (2.3) se u nekim slučajevima (najčešće za određivanje parametra lokacije) može zapisati i kao

$$\sum_{i=1}^{n} w_i(X_i - T_n) = 0, \tag{2.5}$$

gde je T_n implicitno definisan ocenjivač, a w_i su težinski koeficijenti. Ovako zapisan problem je problem težinskih najmanjih kvadrata (nadalje *wLSE*). Ovaj oblik je pogodan za računanje ukoliko je moguće odrediti težine w_i.

2.1 MODEL LOKACIJE

Pretpostavlja se da su $X_1,...,X_n$ *i.i.d.* dati nepoznatom funkcijom raspodele G. Neka je $\{F_\theta : \theta \in \Theta\}$ familija funkcija raspodela gde $\Theta \in R^p$ za neko $p \geq 1$. Cilj je odrediti parametar θ (u opštem slučaju vektor parametara) tako da F_θ daje najbolju aproksimaciju G. Pretpostavlja se $x_i = \theta + u_i$, gde je u_i greška, odakle sledi da je funkcija raspodele grešaka data sa

$$F_\theta(x) = F_0(x - \theta). \tag{2.6}$$

Ovaj model se može proširiti na višedimenzionalni slučaj gde su $X_i, i = 1,...,n$ i θ *q*-vektori za neko $q \geq 1$. Za ocenjivače parametra lokacije u problemu (2.1) i

(2.3) uzimaju se funkcije oblika $\rho(x,\theta) = \rho(x-\theta)$ i $\psi(x,\theta) = \psi(x-\theta)$ i nadalje je pažnja preusmerena samo na ocenjivače parametra lokacije.

Preusmerimo sada pažnju na rešivost problema (2.1) i (2.3).

Teorema 17. *Neka za svako x funkcija $\psi(x-\theta)$ nerastuća funkcija po θ tako da važi*

$$\lim_{\theta\to\theta_1} \psi(x-\theta) < 0 < \lim_{\theta\to\theta_2} \psi(x-\theta).$$

Neka je $g(\theta) = \sum_{i=1}^{n} \psi(x_i-\theta)$. Onda:

- *Postoji bar jedna tačka θ_* u kojoj $g(\theta)$ menja znak.*
- *Za više od jedne takve tačke dobijeni skup je interval.*
- *Ako je ψ neprekidna, onda $g(\theta_*) = 0$.*
- *Ako je ψ opadajuća, onda je tačka θ_* jedinstvena.*

Jedinstvenost može da postoji bez striktne monotonosti ψ. Kao primer može se uzeti Huberova ψ funkcija (data na 36. stranici) koja je nerastuća, ali ocenjeni parametar lokacije je jedinstven pod uslovom da ne postoji veliki raskorak (međuprostor) u sredini podataka (postoje x tako da $|x| < k$).

Neka je ρ parna nenegativna funkcija oblika $\rho(x,\theta) = \rho(x-\theta)$, ovde $\rho(x-\theta)$ meri raskorak između observacije x_i i centra raspodele (tj. nepoznatog parametra) θ, dok $\psi(x-\theta)$ dodeljuje observaciji skor tj. uticaj na ocenjivač. Intuitivno je jasno da će autlajerima biti ograničavan uticaj kroz razne uslove nametnute na ρ i ψ. Dve najinteresantnije klase ψ ocenjivača u zavisnosti od oblika ove funkcije su klasa monotono ograničenih (primer Huberova funkcija) i klasa reopadajućih ocenjivača.

Reopadajuća klasa (redescending class) funkcija je oblika

$$\psi_k(x) = \begin{cases} \psi(x), & |x| \le k \\ 0, & k \le |x|\,. \end{cases}$$

i o njoj će biti posvećena pažnja u narednom delu teksta.

Upravo zbog ovakve kontrole autlajera postavlja se pitanje superiornosti ovih metoda u odnosu na klasične. Ovakvo razmišljanje je pogrešno, jer se opet postavlja pitanje kvantitativne definicije autlajera i u literaturi se takođe mogu naći primeri koji opovrgavaju ovo razmišljanje.

Dakle u zavisnosti od izbora funkcija ρ odnosno ψ dobijaju se različite klase ocenjivača. Primetimo sledeće,

- ukoliko je $\rho(x,\theta) = (x-\theta)^2$, sledi $\psi(x,\theta) = (x-\theta)$ i dobija se metod *LS*[1] metod koji daje uzoračku sredinu $\overline{x_n}$ kao ocenjivač;

[1] Least Squares - Metod namjanjih kvadrata.

- ukoliko je $\rho(x,\theta) = |x-\theta|$, sledi $\psi(x,\theta) = \text{sign}(x-\theta)$ i dobija se *LAV*[2] metod koji daje medijanu $M_e(x)$ kao ocenjivač;
- ukoliko je $\rho(x,\theta) = -\log dF(x)$ onda se očigledno dobija *ML* ocenjivač koji je najbolji metod pod uslovima regularnosti.

Generalno od funkcije ρ se očekuje da bude parna, konveksna funkcija čija je brzina rasta manja nego brzina rasta kvadratne funkcije. Specijalno ako je ρ ograničena funkcija dobija se reopadajuća klasa ψ funkcija.

Monotona klasa je popularna upravo zbog postojanja rešenja problema (2.3) pod poprilično generalnim uslovima od kojih uz neke dodatne uslove dobijaju konzistentni ocenjivači. Navodimo da postoje rešenja ovog problema i konzistentni ocenjivači i za nemonotone funkcije ψ. U narednom delu teksta su navedene neke od ρ i ψ funkcija koje se koriste pri rešavanju problema (2.1) i (2.3) ocenjivača, za sada navodimo još neke teoreme koje garantuju rešenja pomenutih problema.

Definicija 55. Neka je ρ parna, konveksna funkcija tako da važi:

1. ρ je neopadajuća funkcija po $|x|$.
2. $\rho(0) = 0$.
3. ρ je rastuća po $x > 0$ tako da $\rho(x) < \rho(\infty)$.
4. Ukoliko je ρ ograničena, pretpostavlja se da je $\rho(\infty) = 1$.

Funkcije koje zaodovoljavaju navedene osobine zovemo ρ-funkcije.

Definicija 56. Ukoliko je funkcija ψ izvod ρ-funkcije onda će ona biti obeležavana kao ψ-funkcija, i može se pokazati da ona zadovoljava

1. ψ je neparna i $\psi(x) \geq 0$ kada $x \geq 0$.

Za model lokacije rešivosti problema (2.2) i (2.4) zavisi od izbora funkcija ψ i ρ i navodimo par teorema koje garantuju rešivost ovih problema. Problem (2.2), odnosno $E_F[\rho(x-t)]$ ima jedinstven minimum pod uslovima koji važe u Teoremi 18.

Teorema 18. *Neka slučajna promenljiva X ima opadajuću funkciju gustine $f(x)$ po $|x|$, i neka je ρ data ρ-funkcija. Tada $E_F[\rho(x-\theta)]$ ima jedinstven minimum u $\theta = 0$.*

Napomena 1: Ova teorema ne važi za ψ iz reopadajuće klase. Ukoliko je ψ iz reopadajuće klase onda se za postojanje jedinstvenosti mora pretpostaviti i monotonost ψ, kao i da raspodela slučajne promenljive X mora biti simetrična i unimodalna.

Napomena 2: Ukoliko je ψ iz reopadajuće klase i $dF(x)$ nije unimodalna, minimum nije jedinstven.

Teorema 19 (Teorema o monotonoj konvergenciji)**.** *Neka je y_n neopadajući od dole ograničen (odnosno $\exists t \in R$ tako da $\forall n\ y_n > t$) niz slučajnih promenljivih takvih da $E[\![y_n|] < \infty$, i neka $y_n \xrightarrow{s.s.} y$. Tada*

$$E[y_n] \to E[y].$$

[2]Least Absolut Value. U literaturi se koristi i oznaka L_1 za ovu klasu ocenjivača.

Teorema 20. *Neka je* $\mu_F(\theta) = E_F[\psi(x,\theta)]$*, i ako važi* $E_F[|\psi(x,\theta)|] < \infty$ *za svako* $\theta \in \Theta$*. Ukoliko važe uslovi Teoreme 17, onda postoji* θ_* *tako da* $\mu_F(\theta)$ *menja znak u* θ_**.*

Napomena: Ako je μ_F neprekidna, tada $E_F[\psi(x,\theta_*)] = 0$.

Teorema 21. *Ako je* θ_* *jedinstven, onda* $T_n \xrightarrow{p} \theta_*$*.*

Teorema 22. *Ako je* $\psi(x,\theta)$ *monotona po* θ *i* $T(F)$ *je jedinstveno definisana sa* (2.4)*, onda je* T_n *Fišer konzistentna u* F *pa* $T_n \xrightarrow{s.s.} T(F)$ *odakle sledi i* $T_n \xrightarrow{p} T(F)$*.*

Optimalan izbor funkcije zavisi od problema, okruženja i uzorka podataka koji se analizira. O funkcijama i klasama funkcija koje se često koriste detaljnije u ostatku teksta. Detaljnije o uslovima nametnutim na ψ kao i analizi asimptotskih osobina ovih ocenjivača može se pročitati u [20], poglavlje 5.

2.2 Varijansa *M* ocenjivača lokacije

Neka je dat problem (2.4) za datu empirijsku funkciju G_n. Rešenje T_n je zapisano kao $T(G_n)$ gde je T funkcionela data sa

$$\int \psi(x, T(G))dG(x) = 0 \tag{2.7}$$

za raspodelu G za koju je integral definisan. Neka je raspodela G data kao ε-kontaminirano okruženje tj. $G = F_{s,x} = (1-s)F + s\delta_x$. Diferenciranjem problema (2.7) po s dobija se

$$\begin{aligned}\int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)}dF(x) \cdot \frac{\partial}{\partial s}[T(F_{s,x})]_{s=0} \\ + \int \psi(y, T(F))d(\delta_x - F) = 0\end{aligned} \tag{2.8}$$

pod uslovom da se može izvršiti zamena redosleda operacija integracije i diferencijala. Upotrebom (2.7) i definicijom uticajne funkcije (1.5) dobija se uticajna funkcija *M* ocenjivača.

Definicija 57. Uticajna funkcija *M* ocenjivača data je sa

$$IF(x,\psi,F) = \frac{\psi(x,T(F))}{-\int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)}dF(y)}. \tag{2.9}$$

Naravno, uticajna funkcija je dobro definisana pod pretpostavkom da je imenilac različit od nule. Uticajna funkcija data izrazom (2.9) je konačna, odnosno statistika $T(F)$ je *B-robusna* u F ako i samo ako $\psi(x - T(F))$ je ograničena funkcija.

Detaljnije obrazloženje o izrazu (2.9) daje Clarke (1983, 1986). On dokazuje Fréchet diferencijabilnost *M* ocenjivača u generalnom slučaju, čak i kad ψ nije glatka, što dalje implicira postojanje Gáteaux diferencijala čiji je specijalni slučaj upravo uticajna funkcija.

Izrazom 1.15 se dobija i asimptotska varijansa

$$V(T,F) = \int \frac{\psi(x,T(F))^2 \, dF(x)}{[\int \frac{\partial}{\partial\theta}[\psi(y,\theta)]_{T(F)} dF(y)]^2}, \tag{2.10}$$

odnosno

$$V(T,F) = \frac{E_F[\psi^2]}{E_F[(\psi')^2]} \tag{2.11}$$

Ovo je rešenje Huberovog problema za ε-kontaminirani model dat sa (1.9).

2.2.1 Prelomna tačka *M* ocenjivača lokacije

Teorema 23. *Neka je $\psi = \psi(x - T(G))$ ne nužno neprekidna, monotono rastuća funkcija koja menja znak. Onda je* M *ocenjivač T parametra lokacije, definisan sa $\int \psi(x - T(F))dG(x) = 0$ slabo neprekidan u F ako i samo ako je ψ ograničena i $T(F_0)$ je jedinstven ocenjivač. Tačka preloma ε^* je data sa*

$$\varepsilon^* = \frac{\eta}{1+\eta}, \tag{2.12}$$

gde je

$$\eta = \min\Big\{ -\frac{\psi(-\infty)}{\psi(+\infty)}, -\frac{\psi(+\infty)}{\psi(-\infty)} \Big\}. \tag{2.13}$$

Gde tačka preloma dostiže najbolju vrednost $\varepsilon^* = \frac{1}{2}$ ako $\psi(-\infty) = -\psi(+\infty)$.

Tačka preloma je $\varepsilon^* = \frac{1}{2}$ za klasu monotonih i ograničenih ψ funkcijama (poput Huberove funkcije definisane na strani 36). Ova teorema očigledno važi i za reopadajuću klasu ocenjivača.

2.3 Asimptotska efikasnost *M* ocenjivača

Definicija 58. L_p $(1 \leq p < \infty)$ prostor je funkcija za koje je p-ti stepen apsolutne vrednosti ove funkcije Lebeg integrabilan, odnosno

$$\int |f|^p \, dF(x) < \infty.$$

Specijalno, za $p = 2$ dobija se L_2 prostor koji je Hilbertov u smislu da je definisan skalarni proizvod

$$\langle \nu, \xi \rangle = \int \nu(x)\overline{\xi(x)} dF(x) = E_F[\nu\xi].$$

Neka je $\{F_\theta : \theta \in \Theta\}$ familija raspodela i pretpostavimo da je T Fišer konzistentan ocenjivač nepoznatog parametra θ. Pretpostavlja se i da je T Fréchet diferencijablina u funkciji raspodele F_θ.

Neka je f_θ funkcija gustine slučajne promenljive X date familijom raspodela F_θ. Takođe se pretpostavlja da je $d_L(F_\theta, F_{\theta+\delta}) = O(\delta)$, tako da

$$\frac{f_{\theta+\delta} - f_\theta}{\delta f_\theta} \to \frac{\partial}{\partial\theta} log f_\theta \tag{2.14}$$

konvergira u $L_2(F_\theta)$-smislu kada $\delta \to 0$, i važi

$$0 < I(F_\theta) < \infty,$$

onda po definiciji Fréchet diferencijala

$$\begin{aligned} T(F_{\theta+\delta}) - T(F_\theta) - \int IF(x, F_\theta, T)(f_{\theta+\delta} - f_\theta)dx \\ = o(d_L(F_\theta, F_{\theta+\delta})) = o(\delta). \end{aligned} \tag{2.15}$$

Izraz (2.15) se podeli sa δ i pusti da $\delta \to 0$. Budući da je T Fišer konzistentna i imajući u vidu (2.14) sledi

$$\int IF(x, F_\theta, T)\frac{\partial}{\partial\theta}(\log f_\theta) f_\theta = 1. \tag{2.16}$$

Koši-Švarcova nejednakost dalje implicira

$$(\int IF(x, F, T)\frac{\partial}{\partial\theta}(\log f_\theta) f_\theta dx)^2 \leq I(F_\theta) \int IF(x, F_\theta, T)^2 f_\theta dx,$$

odnosno asimptotska varijansa $V(F_\theta, T)$ zadovoljava

$$V(F_\theta, T) = \int IF(x, F_\theta, T)^2 dF_\theta \geq \frac{1}{I(F_\theta)}. \tag{2.17}$$

Dobijeni izraz daje odgovor na pitanje izbora funkcije ψ. Asimptotska efikasnost (odnosno jednakost u izrazu (2.17)) je moguća ukoliko je odabrana funkcija ψ tako da je uticajna funkcija, $IF(x, F, T)$, proporcionalna sa

$$\frac{\partial}{\partial\theta}(\log f_\theta) = \frac{f_0'(x)}{f_0(x)}.$$

Pod navedenim pretpostavkama ocenjivač T je asimptotski efikasan ako i samo ako njegova uticajna kriva zadovoljava

$$IF(x, F_\theta, T) = \frac{1}{I(F_\theta)}\frac{\partial}{\partial\theta}(\log f_\theta). \tag{2.18}$$

Ova jednačina nam daje odgovor na pitanje o optimalnom izboru funkcije ψ kod konkretnog problema.

Za model lokacije gde je $f_\theta(x) = f_0(x - \theta)$ optimalan izbor funkcije je oblika

$$\psi(x) = -c\frac{f_0'(x)}{f_0(x)},\ c \neq 0$$

koja je proporcionalna sa $-\frac{\partial}{\partial\theta}(\log f_0)$. Već je pokazano da ograničena funkcija ψ daje malu asimptotsku varijansu.

Primer 13. *Optimalan izbor funkcije ψ za robusne ocenjivače datih raspodela su*

- *Za logističku raspodelu datu funkcijom raspodele $F_0(x) = \frac{1}{1+e^{-x}}$ optimalan izbor je*
$$\psi(x) = tanh(\frac{x}{2}).$$
- *Za Košijevu raspodelu datu funkcijom gustine $f_0(x) = \frac{1}{\pi(1+x^2)}$ optimalan izbor je*
$$\psi(x) = \frac{2x}{(1+x^2)}.$$

2.4 Asimptotska minimaks teorija

Robusnost i efikasnost metode su dve osobine koje su obrnuto proporcionalne. Postavlja se pitanje koliko se efikasnosti moramo odreći sa ciljem dobijanja robusne metode. Dakle, potrebno je praviti nekakav optimalni kompromis izmedju ove dve osobine. U ovom poglavlju predstavljamo Huberov pristup analiza asimptotskih preformansi (pristrasnosti i varijanse) za rešavanje jednostavnog problema lokacije.

2.4.1 Minimaks pristrasnost

Neka je F prava jednodimenzionalna funkcija raspodele greške koja leži u nekoj okolini $\mathcal{P}_\varepsilon$ pretpostavljenog modela koji ima raspodelu F_0. Observacije su *i.i.d.* sa funkcijom raspodele $F(x-\theta)$ gde je θ nepoznati parametar lokacije koji je potrebno oceniti.

Huber je minimizirao maksimalnu asimptotsku pristrasnost $b(\varepsilon)$ za raspodele $F \in \mathcal{P}_\varepsilon$. Neka je dato okruženje ε-kontaminirane normalne raspodele Φ, odnosno

$$\mathcal{P}_\varepsilon(F_0) = \{F \mid F = (1-\varepsilon)\Phi + \varepsilon H,\ H \in \mathcal{M}\}$$

gde je H iz klase simetričnih, unimodalnih raspodela.

Očigledno maksimum pristrasnosti $b_{max}(\varepsilon)$ medijane se postiže kada postoji jedna izolovana tačka x_0 mase čije je rešenje dato sa

$$(1-\varepsilon)\Phi(x_0) = \frac{1}{2},$$

ili

$$b_{max}(\varepsilon) = x_0 = \Phi^{-1}(\frac{1}{2(1-\varepsilon)}).$$

Potom se konstruiše okolina x_0, odnosno konstruišu se dve ε-kontaminirane raspodele F_+ i F_- koje su simetrične oko x_0 i $-x_0$. Raspodela F_+ data je sa funkcijom gustine

$$f_+(x) = \begin{cases} (1-\varepsilon)\varphi(x), & x \le x_0, \\ (1-\varepsilon)\varphi(x-2x_0), & x > x_0. \end{cases} \tag{2.19}$$

gde je φ funkcija gustine normalne raspodele F. Budući da su $F_+(x)$ i $F_-(x)$ translatorno ekvivalentne, sledi $F_-(x) = F_+(x + 2x_0)$.

Ovo dalje implicira

$$T(F_+) - T(F_-) = 2x_0$$

za svake dve translatorno ekvivalentne raspodele. Ovo pokazuje da ne postoji ocenjivač koji ima apsolutnu pristrasnost manju od x_0 u F_+ i F_- istovremeno.

Zaključujemo da medijana M_e postiže najmanju maksimalnu pristrasnost od svih translatorno ekvivalentnih funkcionela koje su simetrične i unimodalne.

Lema 9. *Za medijanu $T(F) = M_e$ važi $b(\varepsilon) = b_{max}(\varepsilon)$.*

Dokaz. Bez umanjenja opštosti pretpostavlja se da je parametar lokacije raspodele F dat i iznosi $\theta = 0$, odnosno $E_F = 0 = M_e(F)$. Tada je

$$b(\varepsilon) = \lim_{n\to\infty} \sup_{F\in\mathcal{P}_\varepsilon} \left|M_e(T_n, F)\right| \tag{2.20}$$

$$b_{max}(\varepsilon) = \sup_{F\in\mathcal{P}_\varepsilon} \left|T(F_n) - T(F)\right| = \sup_{F\in\mathcal{P}_\varepsilon} \left|T(F_n)\right|, \tag{2.21}$$

pa sledi direktno jer je $T = M_e$.

□

Analogno, za Lévi okruženje (1.8) dobija se isto rešenje. Dakle, medijana rešava minimaks problem minimiziranja maksimalne asimptotske pristrasnosti i ovaj ocenjivač predstavlja najbolji izbor za jako velike uzorke gde je standardna devijacija komparabilna ili manja od pristrasnosti. Analiza maksimalne pristrasnosti je generalno bez previše dešavanja uz napomenu da su korišćeni uslovi simetričnosti i unimodalnosti raspodele, što ne vodi ka preteranoj generalizaciji.

2.4.2 Minimaks varijanse

Rešava se problem minimizacije (2.3), odnosno

$$\sum_{i=1}^{n} \psi_k(x_i - T_n) = 0, \tag{2.22}$$

na svim F iz ε-kontaminirane okoline date sa

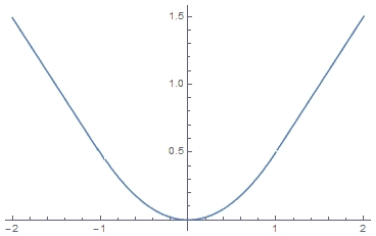
$$\mathcal{P}_\varepsilon = \{F \mid F = (1-\varepsilon)\Phi + \varepsilon H \in \mathcal{M}\},$$

gde je H simetrična, a Φ normalna raspodela. Tejlorov razvoj izraza (2.22) oko T_n daje

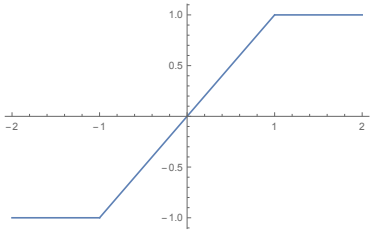
$$\sum \psi_k(x_i) - T_n \sum \psi_k'(x_i) = 0 \tag{2.23}$$

i onda iz centralne granične teoreme sledi asimptotska normalnost $\sqrt{n}T_n$ sa varijansom (2.11), odnosno

$$V(F, T) = \frac{E_F(\psi_k^2)}{(E_F(\psi_k'))^2} \tag{2.24}$$



Slika 2.1: Huberova ρ_k funkcija



Slika 2.2: Huberova ψ_k funkcija

uz komentar da ψ treba da bude ograničena. Huber je definisao funkciju (Slika 2.1)

$$\rho_k(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq k \\ k(|x| - \frac{k}{2}), & k \leq |x| \end{cases},$$

odnosno ψ_k funkciju (Slika 2.2)

$$\psi_k(x) = \begin{cases} x, & |x| \leq k \\ k\,\mathrm{sign}\,(x), & k \leq |x| \end{cases}.$$

Ovo je jedna funkcija iz klase monotonih ograničenih neopadajućih ψ funkcija. Ova funkcija je ograničena pa daje kvantitativne i *B-robusne* ocenjivače. Glavna mana ove funkcije je leži u tome da ona nedovoljno umanjuje uticaj nekim ekstremnim autlajerima, odnosno efikasnost ocenjivača dobijenih ovom funkcojom zavisi od odabira parametra k. Opet napominjemo da je nekada bolje ukloniti ove ekstremne autlajere.

Za Huberovu funkciju analizirana varijansa je

$$V(F,T) \leq \frac{(1-\varepsilon)E_\Phi(\psi_k^2) + \varepsilon k^2}{((1-\varepsilon)E_\Phi(\psi_k'))^2}.$$

Gornja granica ove varijanse dostiže se kada je H ustvari masa u jednoj tački x_0 koja se nalazi van intervala $(-k,k)$.

Ocenjivač definisan sa izrazom (2.22) je *ML* ocenjivač za funkciju

$$f_0(x) = Ce^{-\rho(x)}. \tag{2.25}$$

Pretpostavlja se da je $C = \frac{1-\varepsilon}{\sqrt{2\pi}}$, a parametar k iz Huberove funkcije dat je vezom

$$\frac{2\varphi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon}. \tag{2.26}$$

Odgovarajuća raspodela F_0 je sadržana u $\mathcal{P}_\varepsilon$, i F stavlja svu kontaminaciju u tački van intervala $[-k,k]$, odakle sledi

$$\sup_{F\in\mathcal{P}_\varepsilon} V(F,T) = V(F_0,T).$$

Ocenjivač maksimalne verodostojnosti za F_0 minimizira maksimalnu asimptotsku varijansu za F_0, ali on ustvari minimizira maksimalnu asimptotsku varijansu za sve $F \in \mathcal{P}_\varepsilon$. Ovaj rezultat potrebno je generalizovati za druge okoline $\mathcal{P}_\varepsilon$.

Idealno bi bilo ukoliko bi skup $\mathcal{P}_\varepsilon$ bio kompaktan, ali većina interesantnih okolina $\mathcal{P}_\varepsilon$ nije tesna, pa njihovo zatvaranje nije kompaktno na slaboj topologiji. Zbog ovoga se koristi nejasna (definisana u narednoj definiciji) topologija što kao posledicu ima kompaktnost $\mathcal{P}_\varepsilon$ po cenu uključenja substohastičkih mera u $\mathcal{P}_\varepsilon$. Nadalje se pretpostavlja da je $\mathcal{P}_\varepsilon$ neodređeno zatvoren i odatle kompaktan.

Definicija 59. Nejasna topologija[3] na prostoru $\mathcal{M}+$ substohastičkih mera na Ω je najslabija topologija u odnosu na koju su sva preslikavanja

$$F \mapsto \int \psi dF$$

neprekidna za svaku neprekidnu funkiju ψ sa konačnim nosačem.

Budući da je $\Omega = \mathbb{R}$ onda pored činjenice da je to poljski prostor imamo i osobinu lokalne kompaktnosti, pa odakle sledi da je $\mathcal{M}+$ kompaktnan.

2.4.3 Raspodele koje minimiziraju Fišerovu informaciju

Neka je $F_0 \in \mathcal{P}_\varepsilon$ raspodela koja ima minimalnu matricu Fišerovih informacija među svim raspodelama iz $\mathcal{P}_\varepsilon$ za fiksno ε. Potrebno je pokazati da postoji jedinstvena funkcija raspodele F_0 u okolini $\mathcal{P}_\varepsilon$.

Za dati niz ocenjivača T_n asimptotska varijanasa $\sqrt{n}T_n$ u najboljem slučaju dostiže donju granicu $I(F_0)^{-1}$. Ako je moguće pronaći niz ocenjivača $\{T_n\}$ tako da njegova asimptotska varijansa ostane ograničena sa $I(F_0)^{-1}$ za svako $F \in \mathcal{P}_\varepsilon$ onda je minimaks problem rešen. Ovaj niz mora biti asimptotski efikasan za F_0.

Neka je F_0 raspodela sa najmanjom Fišer informacijom. Prvo će biti proširena definicija Fišerove informacije.

Definicija 60. Fišerova matrica informacija za model lokacije raspodele F definisane na skupu $\mathbb{R}$ data je sa

$$I(F) = \sup_{\psi} \frac{(\int \psi' dF)^2}{\int \psi^2 dF} \tag{2.27}$$

gde $\psi \in C_K^1$ skup svih neprekidno diferencijabilnih funkcija sa kompaktnim nosačem, tako da važi $\int \psi^2 dF > 0$.

Teorema 24. *Sledeći iskazi su ekvivalentni:*

1. $I(F) < \infty$.
2. F *ima apsolutno neprekidnu funkciju gustine* f *i* $\int (\frac{f'}{f})^2 f dx < \infty$.

[3] Vague topology - nejasna topologija.

U bilo kom slučaju važi sledeće $I(F) = \int (\frac{f'}{f})^2 f dx$.

Ako je na $\mathcal{P}_\varepsilon$ definisana nejasna topologija, onda je Fišerova matrica informacija (2.27) semineprekidna sa donje strane kao funkcija od F. Sledi da $I(F)$ dostiže infimum na svakom kompaktnom skupu $\mathcal{P}_\varepsilon$.

Posledica 5 (Egzistencija)**.** *Ako je* $\mathcal{P}$ *neodređeno kompaktan, onda postoji* $F_0 \in \mathcal{P}_\varepsilon$ *koji minimizira* $I(F)$.

Takođe može se pokazati da je $I(F)$ konveksna funkcija od F. Ovo sledi iz činjenice da su $\int \psi' dF$ i $\int \psi^2 dF$ linearne funkcije od F. Dokaz sledi direktno iz leme.

Lema 10. *Neka su* $u(t)$ *i* $v(t)$ *linearne funkcije po* t *tako da* $v(t) > 0$ *za* $0 < t < 1$. *Onda je funkcija* $w(t) = \frac{u(t)^2}{v(t)}$ *konveksna za* $0 < t < 1$.

Posledica 6 (Jedinstvenost)**.** *Pretpostavimo da važi:*

1. $\mathcal{P}_\varepsilon$ *je konveksan.*
2. $F_0 \in \mathcal{P}_\varepsilon$ *minimizira* $I(F)$ *u* $\mathcal{P}_\varepsilon$, *i važi* $0 < I(F_0) < \infty$.
3. *Skup tačaka u kojima je gustina* f_0, *funkcije raspodele* F_0, *striktno pozitivna čini konveksan skup koji sadrži nosač svake raspodele u* $\mathcal{P}_\varepsilon$.

Onda je $F_0 \in \mathcal{P}_\varepsilon$ *jedinstveni minimum koji minimizira* $I(F)$.

Za različite okoline $\mathcal{P}_\varepsilon$ optimalni su različiti izbori funkcija ψ odnosno ρ. Najčešće se koristi klasa monotonih ψ funkcija (npr. Huberova) pri rešavanju problema (2.3).

2.5 Optimalno ograničenje γ^*

U ovom poglavlju su predstavljene ideje o konstrukciji najefikasnijih ocenjivača pod uslovom da gornja granica ukupne osetljivosti γ^* ne prezilazi neki unapred zadati nivo.

Uticajna funkcija, IF, bilo kog autlajera neće prelaziti nivo linearne aproksimirane γ^*. Prvo navodimo jednu teoremu koja će kasnije biti analizirana na *M* ocenjivačima parametra lokacije.

Neka je data funkcija skora maksimalne verodostojnosti

$$s(x, \tilde{\theta}) = \frac{\partial}{\partial \theta} [\ln f_\theta(x)]_{\tilde{\theta}} = \frac{\partial}{\partial \theta} [f_\theta(x)]_{\tilde{\theta}} / [f_{\tilde{\theta}}(x)], \tag{2.28}$$

i neka ona postoji za sve x za koji zadovoljavaju $\int s(x, \theta) dF(x) = 0$[4] i Fišerova matrica informacija zadovoljava $0 < I(F) < \infty$.

[4]Ovaj uslov je potreban za promenu redosleda integracije i diferencijacije u odnosu na θ.

Sada će ova funkcija biti upotrebljena za kreiranje drugih ψ funkcija koja zadovoljavaju (2.4).

Pretpostavlja se da je T Fišer konzistentna $T(F) = \theta_*$. Dalje se pretpostavlja da $f_{\theta_*}(x) \geq 0$ u odnosu na neku meru λ. Diferenciranjem (2.4) po θ za neko fiksno θ_* dobija se još jedan oblik uticajne funkcije

$$\begin{aligned} IF(x, F, T) &= \frac{\psi(x, \theta_*)}{\int \psi(y, \theta_*) \frac{\partial}{\partial \theta}[f_\theta(y)]_{\theta_*} d\lambda(y)} \\ &= \frac{\psi(x, \theta_*)}{\int \psi(y, \theta_*) s(y, \theta_*) dF(y)} \end{aligned} \tag{2.29}$$

u odnosu na meru λ.

Teorema 25 (Hampel, 1968)**.** *Pretpostavlja se da $s(x, \theta_*)$ postoji za svako x i neka važi $\int s(x, \theta_*) dF_{\theta_*} = 0$ i neka Fišerova matrica informacija zadovoljava $0 < I(F_{\theta_*}) < \infty$. Neka je $k > 0$ konstanta. Tada postoji $a \in \mathbb{R}$ tako da*

$$\tilde{\psi}(x) = \begin{cases} -k, & s(x, \theta_*) - a < -k \\ s(x, \theta_*) - a, & \left| s(x, \theta_*) - a \right| \leq k \\ k, & s(x, \theta_*) - a > k \end{cases} \tag{2.30}$$

i zadovoljava $\int \tilde{\psi}(x) dF_{\theta_*} = 0$ i $d := \int \tilde{\psi}(y) s(y, \theta_*) dF_{\theta_*} > 0$. Odnosno, $\tilde{\psi}$ minimizira asimptotsku varijansu

$$V(\psi, F_{\theta_*}) = \int \psi^2 dF_{\theta_*} \Big/ [\int \psi(y) s(y, \theta_*) dF_*(y)]^2 \tag{2.31}$$

među svim preslikavanjima koja zadovoljavaju

$$\int \psi dF_* = 0,$$

$$\int \psi(y) s(y, \theta_*) dF_*(y) \neq 0,$$

$$\sup_x \left| \psi(x) \Big/ \int \psi(y) s(y, \theta_*) dF_*(y) \right| < c, \tag{2.32}$$

gde je

$$c := \frac{k}{d}.$$

Rešenje problema je jedinstveno (do na proizvod sa konstantom).

Dakle za datu klasu ocenjivača definisanu sa (2.22) ukoliko važi prethodna teorema potrebno je primetiti sledeće:

- Uticajna funkcija $IF(x, F, T)$ je definisana izrazom (2.29).
- Ukoliko za trenutak zanemarimo θ_* asimptotska varijansa $V(\psi, F_{\theta_*})$ je definisana sa (2.31) zbog (1.15).

- Rešenje $\tilde{\psi}(\cdot, \theta_*)$ iz Teoreme 25 ima najmanju vrednost varijanse $V(\psi, F_{\theta_*})$ među svim funkcijama $\psi(\cdot, \theta_*)$ koje zadovoljavaju

$$\gamma^*(\psi, F_{\theta_*}) \leq c(\theta_*) = \frac{k(\theta_*)}{d(\theta_*)}.$$

- Ako se ne nametnu gornje granice na γ^* onda se dobijaju ML ocenjivači, odnosno $(b = \infty, a = 0)$. Ukoliko takva funkcija postoji ona je *B-robusna*.

Ova teorema koristi se za kreiranje ψ funkcija i problematika optimalnog izbora ψ funkcije u generalnom slučaju može biti poprilično komplikovana i ona zavisi od osobina koje želimo da nametnemo na ovu funkciju, kao i specifičnosti problema tj. podataka koji se obrađuju. Biće detaljnije i preciznije uvedene dve klase ψ funkcija i ovom problemu neće biti posvećeno više pažnje.

Ovo preslikavanje definiše Fišer konzistentan ocenjivač i za njega postoji IF. Njegova asimptotska varijanasa je minimalna za dato $c = \frac{k}{d}$ na $\gamma^*(\psi, F)$ odakle sledi *B-robusnost* ovog ocenjivača.

Specijalno u slučaju simetričnih funkcija F dobija se $a = 0$, a kada je data $F = \Phi$ onda $\psi_k(x) = x$ što je upravo Huberov ocenjivač. Može se pokazati kada $k \to 0$ dobija se medijana M_e kao ocenjivač parametra.

Kada se radi sa najmanje informativnim raspodelama (npr. (2.25)) one često imaju eksponencijalni rep, odnosno mogu da budu tanjeg repa nego što se može očekivati. Imajući ovo u vidu ponekad je optimalno povećati rizik, tako što ćemo smanjiti uticaj repa raspodele, sa ciljem dobijanja boljih preformansi u nekim raspodelama dugog repa. Kod ovih problema često je u upotrebi reopadajuća klasa ψ-tipa ocenjivača.

2.6 Reopadajuća klasa ψ funkcija

Kao što je pomenuto, ova klasa funkcija ima ograničenu $\rho_k(x)$, odakle sledi

$$\psi_k(x) = \begin{cases} \psi(x), & |x| \leq k \\ 0, & k \leq |x| \,. \end{cases}$$

Ocenjivači koji uzimaju ψ funkciju iz reopadajuća klase su osetljiviji na skaliranje upravo zbog toga što ove funkcije odseku informacije koje se dobijaju iz repa raspodela.

U ovom poglavlju navodimo neke od funkcija koje su imali istorijski značaj i najčešće se javljaju u literaturi. Takođe napominjemo situacije u kojima je optimalno koristiti date ψ funkcije.

- Za datu funkciju raspodele F, sa funkcijom gustine $f(x)$, optimalan izbor funkcije ψ je

$$\psi_F(x) = \begin{cases} -\frac{f'(x)}{f(x)}, & |x| \leq k, \\ 0, & |x| \geq k. \end{cases}$$

- Za ε-kontaminirane normalne raspodele predlaže se rešenje

$$\psi_{a,b,k}(x) = \begin{cases} x, & 0 \le |x| \le a, \\ b\tanh(\frac{1}{2}b(k-x)), & a \le |x| \le k, \\ 0, & k \le |x|\,. \end{cases}$$

Navedena ψ funkcija daje ML ocenjivače za odsečene uzorke sa funkcijom raspodele oblika

$$f_0(x) = \begin{cases} (1-\varepsilon)\varphi(x), & 0 \le |x| \le a, \\ \frac{(1-\varepsilon)\varphi(a)}{\cosh^2(\frac{b}{2}(k-a))}\cosh^2(\frac{b}{2}(k-x)), & a \le |x| \le k, \\ (1-\varepsilon)\varphi(x), & |x| \ge k. \end{cases}$$

Ova funkcija gustine ima prekid u $\pm c$ pa zbog $\int f_0 = 1$ važe uslovi

$$2\int_a^k (f_0(x) - (1-\varepsilon)\varphi(x))dx = \varepsilon,$$

a drugi uslov je $a = b\tanh(\frac{b}{2}(k-a))$.

- Hampelova po delovima linearna ψ funkcija

$$\psi_{a,b,r}(x) = \begin{cases} x, & 0 \le |x| \le a, \\ a\,\mathrm{sign}(x), & a < |x| < b, \\ a\,\mathrm{sign}(x)\frac{(r-|x|)}{r-b}/C, & b \le |x| \le r, \\ 0, & r < |x|\,. \end{cases}$$

gde je $C := \rho(\infty) = \rho(r) = \frac{a}{2}(b-a+r)$.

Ova funkcija se pokazala kao dobra u *"Princeton robusness study"* iz (1972) uz napomenu da ova ψ funkcija ima nagle promene u nagibu, ali zadovoljava Winsorov princip da ψ funkcija koja je linearna u centru donosi veću efikasnost prilikom rada sa normalnim raspodelama.

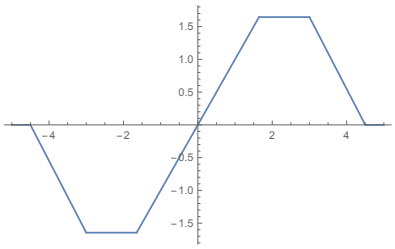
- Takijeva bikvadratna ρ funkcija (ψ glatka i diferencijabilna)

$$\rho_k(x) = \begin{cases} 1 - (1 - (\frac{x}{k})^2)^3, & |x| \le k \\ 1, & k \le |x| \end{cases}$$

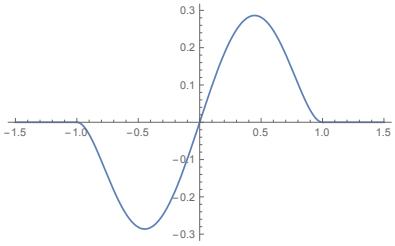
odnosno ψ funkcija

$$\psi_k(x) = \begin{cases} x(1 - (\frac{x}{k})^2)^2, & |x| \le k \\ 0, & k \le |x|\,. \end{cases}$$

Takijeva bikvadratna funkcija sa $k = 4.685$ daje 95% efikasnost na normalnim raspodelama i *M* ocenjivači imaju nisku tačku preloma koja iznosi $\frac{1}{n}$. Ova funkcija se često implementira uz *S* i *MM* ocenjivače.



Slika 2.3: Hampelova funkcija



Slika 2.4: Takijeva bikvadratna funkcija

Danas postoje razne funkcije ψ iz ove klase od koje su prilagođene raznim robusnim kasama ocenjivača (npr. S klasa) i svakim danom se u literaturi mogu naći predlozi novih efikasnih ocenjivača iz ove klase. Na primer:

- Generalizovana Gaus-težinska ψ funkcija

$$\psi_{a,b,k}(x) = \begin{cases} x, & |x| \leq k, \\ x \exp(-\frac{1}{2}\frac{(|x|-k)^b}{a}), & k > |x|\,. \end{cases}$$

- Smitova ψ funkcija

$$\psi_{a,b,k}(x) = \begin{cases} x(k^2 - x^2), & |x| \leq k, \\ 0, & k > |x|\,. \end{cases}$$

Ili ocenjiač priložen u [20],[21], itd. Generalno, funkcije su linearne u sredini, glatke, neprekidne i ne previše strme (u smislu pada i brzine pada) za podatke koji se nalaze dovoljno udaljeno od medijane, ali nedovoljno udaljeno da bi bili autlajeri.

Vrlo grubo govoreći, ispostavlja se da ova klasa ocenjivača ne zavisi preterano od samog oblika ψ funkcije, dok god ψ poštuje neka razumna pravila. Kada biramo funkciju iz ove klase moramo voditi računa o nagibu funkcije ψ i o izvodu funkcije, ψ'. Ona ne sme prebrzo da opada inače nagib može da ima prejak uticaj na imenilac asimptotske varijanase date jednakošću (2.11).

U praksi se, na konačnim uzorcima ova klasa ocenjivača se pokazala generalno efikasnijom od klase monotonih ocenjivača (u nekim slučajevima i do 20% efikasnijom), dok uglavnom beleži gubitke i do 3% eifkasnosti u odnosu na Huberove monotone ocenjivače, što je zanemarljivo.

Huber navodi da je ova klasa delotvorna ukoliko postoje ekstremni autlajeri, ali poboljšanja su minorna i navodi da je uglavnom optimalnije uklanjanje tih par ekstremnih autlajera.

2.7 *M* ocenjivači za parametre skale

Definicija 61. Statistika S_n je ocenjivač skale ukoliko za važi

$$S_n(ax_1, ..., ax_n) = |a| \, S_n(x_1, ..., x_n).$$

Često se zahteva i tranzitivna invarijantnost, ali ova osobina ne mora biti nužno ispunjena za ocenjivače skale.

Definicija 62. Statistika za ocenu skale je tranzitivno invarijantna ako

$$S_n(x_1 + b, ..., x_n + b) = S_n(x_1, ..., x_n).$$

Tri glavna problema su čist problem ocene skale, ocena skale kao pomoćnog parametra i studentizovanje. Čist problem ocene skale se ekstremno retko nalazi u praksi i ovi modeli su uglavnom razvijani zbog teorijskih rezultata, dok se problemi ocene skale kao pomoćnog implementiraju uz ocene regresije ili uz ocenjivače lokacije.

2.8 MODEL SKALE

Model je oblika $x_i = \sigma u_i$ gde su u_i *i.i.d.* sa funkcijom fustine $f(x)$ i gde je σ nepoznat parametar koji se ocenjuje. Slučajna promenljiva X data je familijom raspodela $F(x/\sigma)$, odnosno funkcijom gustine $\frac{1}{\sigma} f(\frac{x}{\sigma})$.

ML ocenjivač nepoznatog parametra σ dat je kao rešenje problema

$$S_n = \arg\max_{\sigma \geq 0} \frac{1}{\sigma^n} \prod_{i=1}^{n} f(\frac{x_i}{\sigma}),$$

odnosno, logaritmovanjem i diferenciranjem po σ dobija se

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\frac{x_i}{\sigma}) = 1$$

gde je $\rho(t) = t\psi(t)$ sa $\psi = -f'/f$.

Primer 14. *Neka je dat prost slučajan uzorak slučajne promenljive X. Standardna (nerobusna) ocena parametra σ se upravo može dobiti iz standardne devijacije. Ocenjivač varijanse dat je sa*

$$s_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

Ovo je Fišer konzistentan ovenjivač varijanse, ali ovaj ocenjivač nije nepristrasan jer

$$E[s_n^2] = \frac{n-1}{n} Var[X]$$

Odavde sledi da je standardna devijacija data sa $\sqrt{s_n^2}$. *Nepristrasan ocenjivač varijanse dat je sa*

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2,$$

ali ovaj ocenjivač varijanse nije Fišer konzistentan.

Definicija 63. *M* ocenjivač skale je statistika $S_n(X_1, ..., X_n) = S_n(F_n)$ i definisana je kao rešenje problema

$$\sigma = \arg\max_{\sigma \geq 0} \frac{1}{n} \sum_{i=1}^{n} \rho(\frac{x_i}{\sigma}) = \delta, \tag{2.33}$$

gde je ρ jedna ρ-funkcija i $0 < \delta < \rho(\infty)$. Ukoliko je ρ ograničena onda se bez umanjenja opštosti pretpostavlja $\rho(\infty) = 1$ i $\delta \in (0, 1)$.

Često se koristi Takijeva bikvadratna ρ funkcija sa $k = 1$. Dovoljan uslov za postojanje jedinstvenog rešenja problema (2.33) tj. ocenjivača skale jeste da je ρ rastuća funkcija po x tako da $\rho(x) < \rho(\infty)$.

Definicija 64. Potpuno ekvivalentno prethodnoj definiciji *M*-ocenjivač za skalu $S(F)$ je definisan implicitno funkcionelom

$$\int \chi\left(\frac{x}{S(F)}\right) dF(x) = 0.$$

Uobičajeno je da je χ parna funkcija. Iz definicije 2.9 uticajna funkcija data je sa

$$IF(x, F, S) = \frac{\chi(\frac{x}{S(F)}) S(F)}{\int \chi'(\frac{x}{S(F)})(\frac{x}{S(F)}) dF(x)}. \tag{2.34}$$

Za ε-kontaminirani model tačka preloma je data sa

$$\varepsilon^* = \frac{-\chi(0)}{\chi(\infty) - \chi(0)} \leq 0.5,$$

za više o tački preloma i modelu čiste ocene skale pogledati [26] poglavlje 5.

U slučaju da problem (2.33) nema rešenje onda $S_n = 0$.

Navodimo jedan od bitnijih pomoćnih ocenjivača skale.

Definicija 65. *MAD*[5] medijana apsolutne devijacije je definisana kao

$$MAD_n = M_e\{|x_i - M_n|\}, \tag{2.35}$$

gde je $M_n = M_e\{x_i\}$, ili preko funkcionele

$$\psi_{MAD}(x) = \text{sign}(|x| - \Phi^{-1}(\frac{3}{4})).$$

[5]Median absolut value.

Ovaj ocenjivač nije Fišer konzistentan, ali se može učititi Fišer konzistentnim. Ukoiko je za $F = \Phi$, onda ga je potrebno pomnožiti sa $\frac{1}{\Phi^{-1}(\frac{3}{4})}$ za Fišer konzistentnost. Za simetrične raspodele *MAD* je asimptotski ekvivalentan polovini međukvartilne distance. Tačka preloma ovog ocenjivača je $\varepsilon^* = 0.5$ i vrlo je robusan u odnosu na autlajere. Za ocenivanje skale kao pomoćnog parametra fokus analize je na repu raspodele.

Primer 15. *Ukoliko je* $\chi(x) = \text{sign}(|x| - 1)$ *tada je* $S = M_e(|x|)$ *tj. broj* S *za koji važi* $F(S) - F(-S) = \frac{1}{2}$.

Problemi ovog oblika su jako retko pojavljuju u praksi i diskusiju nastavljamo o modelima za ocenu više parametara.

2.9 Višeparametarsko ocenjivanje

2.9.1 *M* ocenjivači za parametre skale i lokacije

Budući da *M* ocenjivači lokacije i regresije nisu skalarno invarijantni (uz izuzetak medijane M_e), oni se zato implementiraju uz ocenjivače skale, pa se dobijaju problemi višeparametarskog ocenivanja. Prilikom implementacije mora se obratiti pažnja na tačku preloma ocenjivača, budući da se sad ocenjuju dva ili više parametara. Cilj je izbeći kvarenje dobrih osobina tačke preloma ocenjivača lokacije lošom uslovljenošću tačke preloma ocenjivača skale.

Upravo zbog ovog ocenjivačima parametara lokacije cilj je minimizirati asimptotsku varijansu, dok je za ocenjivače parametara skale poželjno očuvati malu pristrasnost.

Ukoliko je raspodela F modela simetrična, onda su ocenjivači lokacije T_n i skale S_n asimptotski nezavisni i asimptotsko ponašanje T_n zavisi od asimptotskog ponašanja S_n.

Model je oblika $x_i = \theta + \sigma u_i$ gde su u_i dati gustinom raspodele $f(x)$, odakle sledi da slučajna promenljiva X ima gustinu raspodele datu sa

$$f(x) = \frac{1}{\sigma} f(\frac{x - \theta}{\sigma}).$$

Parametar σ je standardna devijacija slučajne promenljive X.

Ocenjivači koji se dobijaju metodom ML dati su sa

$$(T_n, S_n) = \arg\max_{(\theta,\sigma)} \frac{1}{\sigma^n} \prod_{i=1}^{n} f(\frac{x_i - \theta}{\sigma})$$

što se može ekvivalentno zapisati kao

$$(T_n, S_n) = \arg\max_{(\theta,\sigma)} \left[\frac{1}{n} \sum_{i=1}^{n} \rho(\frac{x_i - \theta}{\sigma}) + \log \sigma \right]$$

gde je $\rho = -logf$. Diferenciranjem i izjednačavanjem sa nulom dobijenih jednačina dobijaju se traženi ocenjivači.

Definicija 66. Uporedni *M* ocenjivači za ocenjivanje parametara lokacije $\theta \in \Theta$ i skale σ dati su sistemom jednačina u kom figurišu tražene statistike $T_n = T(F_n)$ i $S_n = S(F_n)$

$$\sum_{i=1}^{n} \psi(\frac{x_i - T_n}{S_n}) = 0 \tag{2.36}$$

$$\sum_{i=1}^{n} \chi(\frac{x_i - T_n}{S_n}) = 0. \tag{2.37}$$

Ekvivalentno ocenjivači $T(F_n)$ i $S(F_n)$ su izražene kao funkcionele

$$\int \psi(\frac{x - T(F)}{S(F)}) dF(x) = 0, \tag{2.38}$$

$$\int \chi(\frac{x - T(F)}{S(F)}) dF(x) = \delta, \tag{2.39}$$

gde je $\delta \geq 0$.

Izbor funkcije ψ, a ni funkcije χ iz prethodne definicije nije determinisan funkcijom raspodele verovatnoće. Kao što je već pomenuto često se za ψ uzima neparna, a χ parna funkcija.

Neka je $\psi_1 = \psi$, a $\psi_2 = \chi$ tada važi sledeća teorema.

Teorema 26. *Neka je za svako x i β funkcija $\psi(x, \beta)$ diferencijabilna tako da je matrica $H = [\frac{\partial \psi_i}{\partial \beta_i}]$ negativno definitna, onda za dato $x_1, ..., x_n$*

$$g(\beta) = \sum_{i=1}^{n} \psi(x_i, \beta).$$

Ukoliko postoji rešenje od $g(\beta) = \mathbf{0}$ onda je ono jedinstveno.

Napomena: Teorema garantuje jedinstveno rešenje problema (2.39) koje je onda potencijalno rešenje problema (2.2).

Uvođenjem $F_t = (1 - t)F + t\delta_x$ umesto F u (2.38) i (2.39) diferenciranjem po t u $t = 0$ mogu se dobiti uticajne funkcije koje zadovoljavaju sistem

$$IF(x, F, T) \int \psi'(y) dF(x) + IF(x, F, S) \int \psi'(y) y dF(x) = \psi(y) S(F), \tag{2.40}$$

$$IF(x, F, T) \int \psi'(y) dF(x) + IF(x, F, S) \int \psi'(y) y dF(x) = \chi(y) S(F), \tag{2.41}$$

gde je $y = [x - T(F)]/S(F)$.

Kada je reč o egzistenciji rešenja sistema datog sa (2.38) i (2.39) o tome se više u [26]. O konzistenciji ovih ocenjivača kao i o njihovoj asimptotskoj normalnosti više se može naći u [26] poglavlju 6.

2.9.2 *M* ocenjivači lokacije sa preleminarnom ocenom skale

Definicija 67. Ukoliko postoji preleminarna ocena skale S_n onda su *M* ocenjivači lokacije dati sa

$$\sum_{i=1}^{n} \psi(\frac{x_i - T_n}{S_n}) = 0.$$

Ovako definisan ocenjivač lokacije je skalarno ekvivalentan. Najčešće se za ocenu skale S_n koristi neka robusna statistika.

Kao robusna alternativa koristi se *MAD* ocenjivač dat u definiciji 65 ili ukoliko je poželjna normalnost ocenjivača, u smislu da ukoliko $X : \mathcal{N}(\mu, \sigma^2)$ za $n \to \infty$ onda $S_n \xrightarrow{p} \sigma$, onda se koristi *MADN*.

Definicija 68. Normalizovana medijana apsolutne devijacije *MADN* je standardizovana medijana apsolutne devijacije, odnosno

$$MADN(x) = \frac{MAD(x)}{\Phi^{-1}(\frac{3}{4})} \approx 1.483 MAD(x).$$

Dakle prvo se oceni

$$S_n = 1.483 MAD(x_i) = 1.483 M_i\{\big|x_i - M_j(x_j)\big|\}$$

gde je M_i medijana i ovaj ocenjivač ima tačku preloma $\varepsilon^* = 0.5$.

U slučaju ocenjivanja lokacije simulacije pokazuju superiornost *M* ocenjivača sa početnom ocenom skale *MAD* (Andrews, 1972) u odnosu na model višeparametarskog ocenivanja skale. Budući da je implementacija modela sa *MAD* takođe značajno jednostavnija, preporučujemo upotrebu ovog modela.

Ukoliko je poznata uticajna funkcija ocenjivača skale, onda je uticajna funkcija ocenjivača lokacije može biti određena iz (2.40).

2.10 Numerička implementacija

U ovom poglavlju demonstriramo neke od prezentovanih modela paketom *MASS* ugrađenim u R-u na konkretno zadatim podacima. U R paketu *MASS* implementiran je Huberov *M* ocenjivač uz *MAD* ocenjivač. Rezultati dati su Tabelom 2.1. Testirani su sledeći ocenjivači

- Lokacije: očekivanje ($\overline{X}$), medijana (M_e), 0.05-trimovano očekivanje ($\overline{X}_{.05}$), 0.10-trimovano očekivanje ($\overline{X}_{.10}$) i Huberov M ocenjivač (M_H).
- Skale: uzoračka standardna devijacija (S_n), interkvartilni opseg (R_I) i medijana apsolutne devijacije (*MAD*).

Primer 16. *Neka je z vektor podataka. 46.34, 50.34, 48.35, 53.74, 52.06, 49.45, 49.90, 51.25, 49.38, 49.31, 50.62, 48.82, 46.90, 49.46, 51.17, 50.36, 52.18, 50.11, 52.49, 48.67.*

Tabela 2.1: Ocenjivači

Podaci	I	II	III	IV	V
$\overline{X}$	50.045	71.946	57.359	0.511	10.993
M_e	50.005	50.225	50.48	-0.040	10.38245
$\overline{X}_{.05}$	50.045	50.327	56.323	-0.0973	10.7594
$\overline{X}_{.10}$	50.089	50.328	55.387	-0.115	10.52294
M_H	50.076	50.328	51.127	-0.101	10.51366
S_n	1.821	97.639	13.665	7.770	3.701513
R_I	2.002	2.09	9.967	1.917	3.394217
MAD	1.742	1.460	2.401	1.440	2.537278

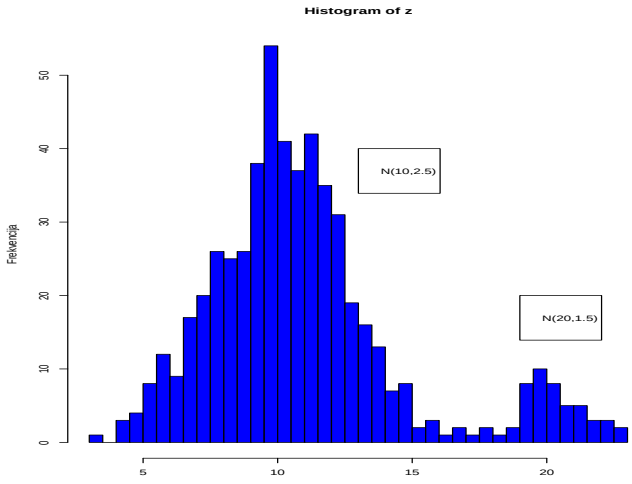
U I koloni su ocenjeni ovi podaci. Za kolonu II simuliran je pogrešan unos podatka i osetljivost statistike na grešku ovog tipa poslednji unos vektora z zamenjen iznosom 486.7.

Za koloni III poslednjih pet vrednosti z zamenjene su redom sa 79.45, 76.80, 80.73, 76.10, 87.01.

Za IV kolonu generisan je i.i.d. *uzorak obima 200 iz Košijeve raspodele date gustinom $f(x) = \frac{1}{\pi(1+x^2)}$ gde je lokacija $\mu = 0$ i skala $\sigma = 1$.*

Za kolonu V smo generisali i.i.d. uzorak obima 500 iz $\mathcal{N}(10, 2.5)$ i i.i.d. uzorak obima 50 iz $\mathcal{N}(20, 1.5)$ i potom smo ta dva uzorka povezali u jedan uzorak.

Dobijeni rezultati dati su u Tabeli 2.1.



Slika 2.5: Histogram podataka simuliran iz dve normalne raspodele

Primetimo da samo jedna loša observacija u podacima vrši znatnu inflaciju klasičnih (nerobusnih) statistika medijane i standardne devijacije, dok je Huberov ocenjivač implementiran sa *MAD* ocenjivačem skale pretrpeo male promene. Slična stvar se događa kada smo u populaciji dobre observacije zamenili lošim observacijama (slučaj III). Primetimo da se ocena skale povećala nekoliko puta. Obratimo pažnju i na slučaj V kada 10% podataka dolazi iz kompletno druge raspodele, Huberov ocenjivač je uspeo da preciznije oceni tražene parametre.

Glava 3

Robusna regresija

Generalno problem linearne regresije je problem oblika

$$\hat{Y} = \arg\min_{\hat{Y}} \left\| Y - \hat{Y} \right\|,$$

gde je Y predstavlja registrovane, a $\hat{Y}$ fitovane vrednosti. Oblik veze između nepoznatih parametara modela, β, koje je potrebno oceniti i observacija X je $\hat{Y} = \beta_0 + \beta_1 X_1 +, ..., +\beta_n X_n$. Za različit izbor norme dobijaju se različiti ocenjivači. Problem se uopštava uvođenjem ρ i ψ funkcija koje umesto norme sada mere rastojanje $Y - \hat{Y}$. Postoji mnogo pristupa problematici robusne regresije, ali istorijski većina metoda su fokusirane na očuvanje visoke tačke preloma (videćemo da i *M* ocenjivači imaju tačku preloma $\varepsilon^* = 0$), dok je kriterijum efikasnosti bio znatno zanemaren. Stoga prvo uvodimo metodu klasičnih *M* ocenjivača, a potom i neke njene podklase. Situacija se menja kada Yohai (1987) uvodi klasu *MM* ocenjivača koja kombinuje visoku relativnu efikasnost *M* ocenjivača sa osobinom visoke tačke preloma *S* ocenjivača.

Posle kratkog osvrta na klasične (nerobusne) modele linearne regresije, odnosno metode *LS* i *ML* ocenjivača, predstavljamo i metod težinskih najmanjih kvadrata (nadalje *wLSE*). Sistem *M* ocenjivača regresije implementiran je preko sistema težinskih najmanjih kvadrata preko težinskih funkcija, w, koje zavise od izbora funkcija ψ, odnosno ρ. Takođe, u ovoj glavi deo pažnje posvećen je klasičnim metodama za detekciju autlajera i analizu uticajnih tačaka i ovi rezultati su kratko i precizno definisani uglavnom bez detaljnog izvođenja.

Potom je pažnja posvećena modelu višeparametarske regresija *M* ocenjivača, nekim njegovim prednostima i manama. Dalje po detekciji nekih mana ove klase, uvodimo još par robusnijih podklasa *M* ocenjivača, odnosno *GM* i *MM* ocenjivače regresije. Potom navodimo i analiziramo uticajne funkcije i relevantne funkcionele ovih ocenjivača. U poslednjem poglavlju ilustrujemo implementaciju kroz statističke pakete u R-u na konkretnim primerima.

Prilikom prezentovanja klasičnih metoda pažnja neće biti posvećena detekciji multikolinearnosti, heteroskedastičnosti ili odstupanju od normalnosti raspodele, jer upravo se robusna regresija uvodi iz potrebe za metodama koje su neosetljive na devijacije od standardnih pretpostavki modela koje su često pogodna matematička

racionalizacija stvarnosti. Takođe kada je reč o robusnom testiranju hipoteza, robusnim matricama kovarijanse i korelacije i intervalima poverenja detaljnije se može naći u literaturi [13], [26], [29] i drugoj literaturi koja se bavi ovom problematikom.

3.1 Klasični modeli linearne regresije

3.1.1 Metode *LS* i *ML*

Model linearne regresije sa p nepoznatih parametara pretpostavlja vezu oblika

$$Y = X\beta + \epsilon,$$

gde je

$Y = [Y_1, ..., Y_n]^T_{n\times 1}$ vektor zavisnih (u funkcionalnom smislu) vrednosti,

$X = [x_{ij}]_{n\times(p+1)}$ je nesingularna matrica observacija (nezavisnih vrednosti),

$\beta = [\beta_0, ..., \beta_p]^T_{(p+1)\times 1}$ vektor nepoznatih parametara iz otvorenog konveksnog skupa $\mathcal{B} \subset R^{p+1}, (p+1) \leq n$

$\epsilon = [\epsilon_1, ..., \epsilon_n]^T_{n\times 1}$ je vektor slučajnih promenljivih koji predstavlja greške ili slučajno odstupanje sa $E[\epsilon] = 0$ i $Cov[\epsilon, \epsilon'] = \Gamma$.

(A) Pretpostavlja se da važi:

1. $E[\epsilon] = 0$.
2. $Cov[\epsilon] = \sigma^2 I_n$ gde je σ nepoznata varijansa.

Definicija 69. Vektor reziduala definisan je kao $r = Y - \hat{Y} = Y - X\hat{\beta}$. Nadalje će i-ti rezidual biti obeležen sa $r_i(\hat{\beta})$.

Ukoliko važi pretpostavka (A) metodom *LS* dobija se ocenjivač B vektora nepoznatih parametara β tako da

$$B = \arg \min_{B\in R^{p+1}} \|r\|_2 . \tag{3.1}$$

Diferenciranjem (3.1) dolazi se do sistema normalnih jednačina

$$X^T X B = X^T Y$$

i ocenjivača *LS*

$$B = (X^T X)^{-1} X^T Y.$$

Pod navedenom pretpostavkama može se pokazati Fišer konzistentnost i nepristrasnost ovog ocenjivača. Takođe može se pokazati $Cov[B, B^T] = \sigma^2 (X^T X)^{-1}$, drugim rečima metoda *LS* je *BLUE* (Teorema Gauss-Markov). Ocenjivač nepoznatog parametra σ^2 dat je sa

$$s^2 = \frac{r^T r}{n - p - 1}.$$

(B) Pretpostavlja se da važi i

3. $\epsilon \sim N_n(0, \sigma^2 I_n)$, tj. normalnosti slučajne promenljive ϵ.

U prisustvu odstupanja od normalnosti vektora grešaka $\epsilon \sim N_n(0, \sigma^2 I_n)$ metoda *LS* za ocenu parametra B drastično gubi efikasnost, što nije poželjna osobina.

Navedene pretpostavke (A) i (B) čine klasične pretpostavke pod kojim je razvijena klasična teorija regresije. Pod ovim pretpostavkama razvija se *ML* metod. Ukoliko važe navedene pretpostavke, može se pokazati sledeće:

1. B je *ML* ocenjivač nepoznatog parametra β (nadalje označavan sa $\hat{\beta}$).
2. $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2 (X^T X)^{-1})$.
3. $(n-p-1)s^2/\sigma^2 \sim \chi^2_{n-p-1}$
4. Standardna greška $\hat{\beta}_i$ data je kao $SE[\hat{\beta}_i] = s\sqrt{[(X^T X)^{-1}]_{ii}}$.

Fitovane vrednosti obeležavaju se sa $\hat{Y} = X\hat{\beta}$ što je projekcija Y na prostor potporostor generisan vektor kolonama nesingularne matrice X. Fitovane vrednosti zadovoljavaju $\hat{Y} = HY$, gde je $H = X(X^T X)^{-1} X^T$. Osobine i struktura matrice H, posebno dijagonale karakteriše interpolaciju. Kvalitetna analiza ove matrice obezbeđuje dijagnostikovanje autlajera, uticajnih tačaka i devijacije od pretpostavki modela.

Kolona $H_j = [h_{1j}, ..., h_{nj}]^T$ matrice H zove se j-ti leveridž vektor i pokazuje kako se $\hat{Y}_j$ menja u odnosu na j-tu posmatranu varijablu, dok $\|H_j\|_2^2$ predstavlja j-ti leveridž, odnosno meru ukupnog uticaja j-te observacije na fitovanu vrednost. Može se pokazati $\|H_j\|_2^2 = h_{jj}$, odakle sledi da je leveridž j-te observacije upravo h_{jj}.

Osnovne osobine matrice H:

- Matrica H je (ortogonalna) projekcija iz R^n na range generisan vektor kolonama matrice X. Za matrice projektora važi $H = H^T$ i $H = H^2$.
- Leveridž x_i^T dat je sa $h_{ii} = x_i^T (X^T X)^{-1} x_i$.
- Iz idempotentnosti matrice H važi

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geq h_{ii}^2,$$

  što implicira da dijagonalni elementi matrice $h_{ii} \in [0, 1]$.

- $tr(H) = \sum h_{ii} = p + 1$, što implicira da je prosečan leveridž $\frac{1}{n}\sum h_{ii} = \frac{p+1}{n}$.

Tačke sa većim h_{ii} su po definiciji tačke jakog uticaja. Šta se tačno smatra značajnim je diskutabilno. Jedan od predloga je da se tačke za koje $h_{ii} \geq \frac{p+1}{n}$ smatraju leverage tačkama. U praksi se često nailazi da vrednosti $h_{ii} \leq 0.2$ smatraju bezbednim, tačke $0.2 \leq h_{ii} \leq 0.5$ smatraju rizičnim i kada $h_{ii} \geq 0.5$ preporučljivo je

izostavljanje ovih tačaka iz regresije. Jedan od pristupa robusifikaciji ovih metoda može biti upravo kroz nametanje uslova na observacije kroz matricu H mora da zadovoljava i njenu dekompoziciju. Ovo su takozvani ekvileveridž dizajnovi. Za detaljnije pogledati [23].

Posledica 7. *Pretpostavimo da važi* $E[\epsilon] = 0$ *i* $Cov[\epsilon, \epsilon^T] = \sigma^2 I_n < \infty$. *Tada je* $\hat{Y}_i$ *konzistentan ako i samo ako* $h_{ii} \to 0$, *a vektor* $\hat{Y}$ *je konzistentan ako i samo ako*

$$h = \max_{1 \leq < n} h_{ii} \xrightarrow{n \to \infty} 0.$$

Teorema 27. *Za svako* $N = 1, 2, ...$ *pretpostavimo da je regresiona jednačina data sa*

$$Y_N = X_N \beta_N + \epsilon_N$$

gde je X_N *dizajn matrica dimenzija* $n_N \times p_N$, *vektor grešaka je* ϵ_N i.i.d. *i važi* $E[\epsilon] = 0$ *i* $Var[\epsilon] = \sigma^2$. *Ukoliko* $H_N = X_N (X_N^T X_N)^{-1} X_N^T$ *zadovoljava*

$$\max_{1 \leq i \leq n_N} h_{ii} \to 0, \textit{kada } \; n_N \to \infty.$$

Onda je $\hat{\beta_N}$ *asimptotski normalan ocenjivač parametra* β.

Napominjemo da navedena teorema ne zahteva normanost vektora ϵ. Navedene teoreme važe u slučaju nestohastičkih objašnjavajućih varijabli. Kada su objašnjavajuće varijable stohastičke onda se situacija komplikuje.

3.1.2 Težinski najmanji kvadrati

Specijalno pretpostavimo da važi $Cov[\epsilon, \epsilon^T] = \sigma^2 W^{-1}$ gde je W dijagonalna, pozitivno-definitna matrica sa poznatim težinama w_i na dijagonali. Tada je problem težinskih najmanjih kvadrata dat sa

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \sum_{i=1}^{n} w_i r_i^2 = \arg\min_{\beta \in \mathcal{B}} \; r^T W r. \tag{3.2}$$

Rešenje ovog problema daje ocenjivač $\hat{\beta}$ koji navodimo u narednoj teoremi.

Teorema 28. *Najbolji linearan nepristrasan ocenjivač (BLUE) parametra* β *je dat sa*

$$\hat{\beta} = (X^T W X)^{-1} X^T W Y,$$
$$Cov[\hat{\beta}, \hat{\beta}^T] = (X^T W X)^{-1}.$$

U opštem slučaju $Cov[\epsilon, \epsilon'] = \Gamma$ ocenjivač parametra β dat je sa

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \sum_{i,j=1}^{n} r_i \Gamma_{ij} r_j = \arg\min_{\beta \in \mathcal{B}} r^T \Gamma r.$$

Biranje adekvatne težinske funkcije $w(r)$ koja se koristi za skaliranje reziduala. Bez obzira na izbor težinske funkcije w cilj je težinskim koeficijentima $w_i(r_i)$ smanjiti uticaj velikih reziduala, odnosno vrednosti koje su udaljene od klastera podataka, oslanjajući se na pretpostavku o nezavisnosti vektora grešaka.

Tradicionalno, *wLSE* metoda se koristi za probleme gde se javlja heteroskedastičnost. Ova metoda se implementira iterativno (IRWLS algoritmom) u svakom koraku ažurirajući težinske koeficijente na osnovu reziduala. Po samoj konstrukciji ova metoda je robusnija na autlajere od navedenih klasičnih metoda i zbog njene jednostavnosti i elegancije često je u upotrebi.

3.1.3 Reziduali i autlajeri

U ovom paragrafu navode se neke veze koje važe između observacija i fitovanih vrednosti. Pažnja je uglavnom posvećena analizi reziduala, uticajnih tačaka i autlajera i načinu na koji utiču na regresiju.

Vektor reziduala dat je izrazima

$$r = (I - H)Y = (I - H)\epsilon,$$

odnosno,

$$r_i = (1 - h_{ii})\epsilon_i - \sum_{j \neq i} h_{ij}\epsilon_j.$$

Potrebno je uočiti da ukoliko $h_{ii} \approx 1$, onda moguća neočekivana greška u Y_i nije nužno odražena u rezidualu r_i, odnosno dolazi do maskiranja grešaka. Ovo znači da veličina reziduala r_i ne mora biti povezana sa sa greškom u observaciji Y_i, već sa nekom drugom observacijom koja ima veliku vrednost h_{ij}.

Posmatrajući izraze $Var[\hat{Y}_i]$ i $Var[r_i]$, odnosno dijagonalu $Cov[\hat{Y}] = \sigma^2 H$ i $Cov[r] = \sigma^2(I - H)$ vidi se da velikim leveridžom imaju za posledicu fitovanu vrednost sa velikom varijansom i odgovarajući rezidual sa malom varijansom.

Da bi se reziduali međusobno poredili potrebno ih je standardizovati. Neka oznaka (i) u indeksu $(X_{(i)}, Y_{(i)})$ označava skup podataka gde su (x_i^T, Y_i) izostavljene iz uzorka. Tada je ocena parametra β data sa $\hat{\beta}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)} Y_{(i)}$. Onda je moguće izračunati $\hat{Y}_{(i)} = x_{(i)}^T \hat{\beta}_{(i)}$ (u x_i^T) baziranu na podacima koji su ostali. Potom se definiše i-ti predviđeni rezidual kao $r_{(i)} = Y_i - \hat{Y}_{(i)}$.

Definicija 70. Studentizovani rezidual definisan je kao

$$r_{(i)}^* = \frac{Y_i - \hat{Y}_{(i)}}{\widehat{SE}[Y_i - Y_{(i)}^*]},$$

gde je $\widehat{SE}$ odgovarajuća ocena standardne greške.

Pod klasičnim pretpostavkama važi sledeće

$$r_{(i)} = Y_i - \hat{Y}_{(i)} = \frac{Y_i - Y_{(i)}}{1 - h_{ii}} = \frac{r_i}{1 - h_{ii}},$$

i važi

$$Var[Y_i - Y_{(i)}] = \frac{\sigma^2}{1 - h_{ii}}.$$

Ovaj poslednji rezultat pokazuje da je $SE[Y_i - \hat{Y}_{(i)}] = \sigma/\sqrt{1-h_{ii}}$, gde se parametar σ ocenjuje statistikom za standardnu grešku $s_{(i)}$ i može se lako izračunati iz veze $(n-p-2)^2 s^2_{(i)} = (n-p-1)s^2 - \frac{r_i^2}{1-h_{(ii)}}$.

Definicija 71. Studentizovani rezidual definisan je kao

$$r^*_{(i)} = \frac{r_i}{s_{(i)}\sqrt{(1-h_{(ii)})}}.$$

Ako važi pretpostavka o normalnosti onda raspodela $r^*_{(i)} \sim t_{n-p-2}$. Studentizovani reziduali nisu međusono nezavisni pa odbacivanje observacije (x_i^T, Y_i) kada $\left|r^*_{(i)}\right| \geq t_{n-p-2}(1-\alpha/2)$ za neko α ne nosi nivo poverenja kada se posmatraju više ovakvih implikacija, odnosno statistiku $\left|r^*_{(i)}\right|$ više treba posmatrati kao indikator na moguće komplikacije u uzorku.

Sada će pažnja biti preusmerena na efekte koji observacija (x_i^T, Y_i) ima na fitovanu vrednost. Neka je matrica X dobijena permutacijom vrsta tako da

$$X = \begin{bmatrix} X_{(i)} \\ x_i^T \end{bmatrix}$$

odakle sledi $X^T_{(i)}X_{(i)} = X^TX - x_i x_i^T$. Uticaj i-te observacije (x_i^T, Y_i) na $\hat{\beta}$ i $\hat{Y}_i$ dat je kroz relacije

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X^TX)^{-1}x_i}{1-h_{ii}} r_i,$$

$$\hat{Y}_i - \hat{Y}_{(i)} = \frac{h_{ii}}{1-h_{ii}} r_i,$$

respektivno. Iz ovih relacija može se uočiti ako je $h_{ii} \approx 1$ dolazi do znatne inflacije ovih pokazatelja.

Kuk[1] predlaže da se uticaj i-te observacije meri sa

$$D_i = \frac{(\hat{Y} - \hat{Y(i)}))'(\hat{Y} - \hat{Y}(i))}{ps^2},$$

odnosno,

$$D_i = \frac{1}{p}(r_i^*)^2 \frac{h_{ii}}{1-h_{ii}}.$$

Vrednosti $D_i \approx 1$ se smatraju velikim i potrebno ih je dalje ispitati.

Budući da se $SE[\hat{Y}_i] = \sigma\sqrt{h_{ii}} \approx s_{(i)}\sqrt{h_{ii}}$ može se meriti standardizovana razlika fitovanih vrednosti

$$DIFTS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{s_{(i)}\sqrt{h_{ii}}} = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{\frac{1}{2}} r^*_{(i)}.$$

[1] Cook's distance - Kukova distanca.

Predlaže se dalja analiza validnosti observacije kada je

$$|DIFTS_i| > 2\sqrt{\frac{p+1}{n-p-1}}.$$

U praksi se često za dijagnostiku umesto koeficijenta 2 u prethodnoj statistici koristi koeficijenat 1.5.

Još neke osobine matrice H:

5. Ukoliko je matrica X slabo uslovnjena u smislu da su dve kolone približno paralelne, onda postoji kolinearnost (zavisnost) u uzorku. Nije moguće izvršiti nikakvu transformaciju matrice H tako da se problem kolinearnosti otkloni. U ovom slučaju predlaže se izbacivanje jedne od observacionih promenljivih.

6. Moguće je smanjiti leverage i-te tačke bez uklanjanja iste pod uslovom da se još observacija uzme iz prostora koji generiše matrica X.

Kao što je već rečeno mala devijacija od standardnih pretpostavki modela utiče na klasične metode ocenjivača $\hat{\beta}$. Ukoliko je poznato da je narušena neka od osnovnih pretpostavki (npr. homoskedastičnost ili očekivanje) postoje razvijene adekvatne metode ili rešenja koja se bave ovom problematikom.

Idealna robusna metoda za ocenjivač parametra β treba da poseduje osobine:

- Konzistentnosti, asimptotsku normalnosti i visok stepen efikasanost pod pretpostavkom da klasične pretpostavke nisu narušene.
- Neosetljivosti na malu devijaciju u pretpostavkama modela.
- Takođe, poželjna je i jednostavnost modela u smislu komputacije.

3.2 Slučajne objašnjavajuće varijable

Postalja se pitanje šta se dešava kada su objašnjavajuće observacije x takođe slučajnih promenljive. Pretpostavlja se da vrste matrice X dolaze iz p-dimenzionalne raspodele. Neka je G zajednička raspodela $(p+1)$-dimenzionalnog vektora $[x^T, y]$, gde je x_i^T slučajna vrsta iz X, a y_i je odgovarajuća slučajna varijabla koja odgovara ovoj vrsti. Razlog zašto x posmatramo kao slučajnu promenljivu jeste upravo zbog mogućnosti modeliranja potencijalnih grešaka u matrici X. Drugi razlog zašto ovo radimo je upravo zbog posmatranja modela kroz već uvedenu aparaturu funkcionela i funkcija uticaja. Koje su to pretpostavke potrebne da se problem regresije zapiše u standardnom obliku $y = x^T\beta + \epsilon$?

Intuitivno je potrebno razmišljati o slučajnoj promenljivoj x kao o marginalnoj, a o promenljivoj y razmišljati kao o uslovnoj raspodeli $y|x$ datoj nekom funkcijom raspodele $F_{y|x}$. Pod pretpostavkom da su $y - x^T\beta$ i $x^T\beta$ nezavisni i imajući u vidu ovaj način razmišljanja problem regresije je dobro definisan i može se odrediti raspodela slučajne promenljive ϵ preko

$$y - x^T\beta = \epsilon \sim F_\epsilon.$$

Pretpostavka o nezavisnosti je potrebna da bi odgovarajući problem regresije bio definisan.

3.3 Linearna regersija *M* ocenjivača

U proteklih 40. godina razvio se veliki broj robusnih ocenjivača parametra β. Većina ocenjivača, ili imaju problem manje relativne efikasnosti (u odnosu na klasične metode), ili imaju probleme sa prelomnom tačkom. Prvo uvodimo klasične *M* ocenjivače, a zatim će pažnja biti posvećena nekim podklasama ove klase ocenjivača.

Definicija 72. ρ - tip *M* ocenjivača parametra $\beta \in \mathcal{B}$ predstavlja rešenje problema

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \rho(y_i - x_i^T \beta). \tag{3.3}$$

Ovde se, kao i u slučaju modela lokacije, sasvim prirodno nameće uslov (stroge) konveksnosti funkcije ρ upravo zbog postojanja (jedinstvenog) minimuma ovog problema i asimptotskih osobina ocenjivača. Ukoliko je ρ konveksna funkcija, onda se problem (3.3) može ekvivalentno zapisati u obliku navedenom u narednoj definiciji. Iz konveksnosti sledi konzistentnost rešenja (3.4) i asimptotska jedinstvenost ocenjivača $\hat{\beta}$. Dakle ako je ρ konveksna funkcija dobija se jedan problem NLP[2] bez ograničenja budući da $\beta \in R^{p+1}$. Pominjemo samo da postoje klase ocenjivača koje rešavaju ovaj problem nekom od metoda nelinearnog programiranje (npr. Njutn-Rapsonova metoda).

Ipak, rešenje ovog problema postoji i za funkcije ρ koje nisu nužno (strogo) konveksne. Detaljnije o konzistentnosti ρ klase ocenjivača u [41].

Definicija 73. ψ - tip *M* ocenjivača parametra $\beta \in \mathcal{B}$ predstavlja rešenje problema

$$\sum_{i=1}^{n} \psi(y_i - x_i^T \hat{\beta}) x_i = \mathbf{0}, \tag{3.4}$$

gde je $\psi = \frac{\partial}{\partial \beta} \rho$.

Generalno, kao i u slučaju modela lokacije, rešavanje problema (3.3) ili (3.4) zavisi od izbora ψ, odnosno ρ. Ukoliko ρ nije konveksna, pomenuti problemi nisu ekvivalentni, odnosno nemaju nužno isti skup rešenja. Sistem (3.4) neretko ima više rešenja (relativnih minimuma) od kojih je samo jedno rešenje globalni minimum 3.3.

Takođe interesuje nas rešenje problema (3.4) za funkcije ψ i pod kojim uslovima su ova rešenja zapravo rešenja problema (3.3). Detaljnije o ovom problemu i metodama kojima se broj rešenja problema (3.4) smanjuje može se pročitati u [28]. Najčešće se pretpostavlja da je ψ bar neprekidna i ograničena funkcija.

Upravo zbog uticaja funkcije ψ na varijansu u fokusu ovog rada pažnja je isključivo posvećena klasama monotonih i reopadajućih funkcija ψ odakle se dobijaju monotoni i reopadajući *M* ocenjivači. Glavna prednost monotonih ψ ocenjivača je

[2]Nonlinear programing.

upravo u tome da su sva rešenja (3.4) upravo rešenja (3.3). Nadalje ako je ψ rastuća funkcija onda je rešenje jedinstveno.

Zbog jednostavnosti sistema u praksi se najčešće rešava problem (3.4). U [20], poglavlju 5, može se naći dokaz za prilično široku klasu funkcija ψ za koje postoji bar jedno rešenje problema (3.4) koje je $\sqrt{n}$-konzistentan ocenjivač nepoznatog parametra β. U [26] poglavlju 7, postoji detaljnija diskusija o uslovima i konvergenciji rešenja (3.4) ka (3.3).

Ovako definisani *M* ocenjivači nepoznatog parametra β, ipak nisu robusni na same greške u matrici nosača X ili u slučaju kada je x slučajna promenljiva. U slučaju kada postoje tačke sa velikim leveridžom, onda ocenjivači parametra β mogu biti neopuzdani, što se i direktno vidi iz izraza (3.4). Dakle, tačka preloma klasičnih *M* ocenjivača je $\varepsilon^* = 0$ kada je reč o tačkama sa velikim leveridžom. Ovo se može videti sa Slike 3.1.

Huber navodi da tačke sa jakim uticajem (velikim leveridžom) potrebno identifikovati i ukoliko je potrebno ukloniti, radije nego celokupan proces rešavanja problema ocene parametra prepustiti robusnim metodama.

Autlajeri ne moraju nužno biti *loše* observacije i ponekada uklanjanje ovih tačaka nije opravdano. Generalno uočavanje ovih vrednosti u slučaju višeparametarskih skupova podataka predstavlja problem. Upravo u ovim slučajevima optimalno je koristiti robusne regresione metode.

Nadalje pretpostavlja se da $h = \max h_{ii} \ll 1$. Grubo govoreći u ovom slučaju se može reći da matrica nosača $X = [x_{ij}]$ ne sadrži ogromne greške, koje bi znatno uticale na rezultate ovih ocenjivača.

Generalno klasični *M* ocenjivači nisu skalarno ekvivalentni, ali se taj problem rešava, kao u slučaju lokacije. Skalarno ekvivalentni *M* ocenjivači dati su sa

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \rho(\frac{r_i(\beta)}{S_n}). \tag{3.5}$$

gde je S_n parametar skale koji se mora oceniti. Diferenciranjem (3.5) dobija se sistem

$$\sum_{i=1}^{n} \psi(\frac{r_i(\hat{\beta})}{S_n}) x_i = \mathbf{0}, \tag{3.6}$$

$$\frac{1}{n} \sum_{i=1}^{n} \chi(\frac{r_i(\hat{\beta})}{S_n}) = \delta. \tag{3.7}$$

Često je χ ustvari jedna ρ-funkcija. Upravo ovo će omogućiti povećanje prelomne tačke statistike bez preternog gubitka efikasnosti. Kasnije navodimo klasu *MM* ocenjivača. Ovo je problem višeparametarskog ocenivanja. Za razliku od problema lokacije u ovom slučaju nije preporučljiva procena parametra skale unapred. Napominjemo da nije moguće oceniti skalu, S_n, bez da se prvo oceni parametar $\hat{\beta}$, koji sam po sebi zahteva ocenu skale.

Navodimo samo da se *M* ocenjivači definisani sa (3.6) u slučaju glatkih funkcija ψ mogu zapisati u obliku ocenjivača težinskih najmanjih kvadrata, odnosno

$$\sum_{i=1}^{n} w_i(y_i - x_i\hat{\beta})x_i = \mathbf{0} \tag{3.8}$$

gde je $w(r) = \frac{\psi(r)}{r}$, pa se za izračunavanje koristiti IRWLS[3] algoritmi uz uslove

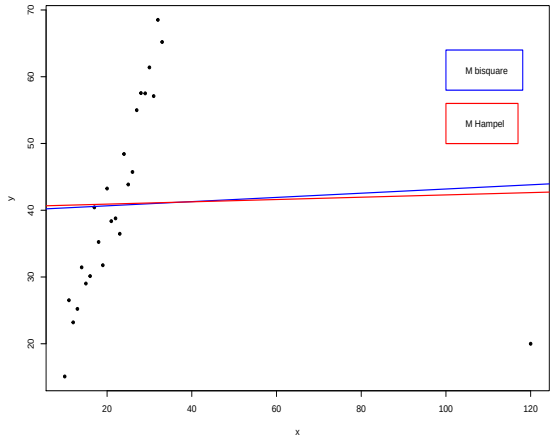
$$\frac{\left\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\right\|}{\left\|\hat{\beta}^{(k+1)}\right\|} < \tau,$$

ili

$$\frac{\left\|r^{(k+1)} - r^{(k)}\right\|}{\left\|r^{(k+1)}\right\|} < \tau,$$

gde je $\tau = 10^{-4}$.

Odabir funkcije $w(r)$, koja se koristi za skaliranje reziduala, direktno je povezana sa biranjem funkcije raspodele koju imaju greške ϵ. Ipak, izborom funkcije w koja određuje *robusan* ocenjivač, nije neophodno da greške nužno prate baš tu raspodelu.



Slika 3.1: Uticaj tačaka leveridža na klasične *M* ocenjivače.

3.3.1 M ocenjivači sa ograničenom ρ funkcijom

Ukoliko je ρ diferencijabilna funkcija, dobija se problem (3.6), odnosno

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\beta})}{S_n}\right)x_i = \mathbf{0},$$

gde je ψ reopadajuća funkcija. Kao posledica, može se dobiti više rešenja (lokalnih minimuma) ovog problema od kojih je samo jedno rešenje globalno rešenje problema (3.5).

[3]iterative reweighted least squares. IRWLS algoritam je moguće naći u Draper i Smith (1998) str. 572.

Definicija 74. Statistika T_n je regresiono invarijantna ako za $\beta \in R^p$ važi

$$T_n(X, y + X\beta) = T_n(X, y).$$

Definicija 75. Statistika T_n je afino invarijantna ako za nesingularnu matricu $A \in R^{p\times p}$ važi

$$T_n(XA, y) = T_n(X, y).$$

Ukoliko je ocenjivač skale S_n regresiono i afino invarijantan i skalarno ekvivalentan onda i ocenjivač $\hat{\beta}$ definisan sa (3.5) poseduje iste ove osobine.

Može se pokazati da je prelomna tačka ovih ocenjivača na konačnom uzorku ocenjivača sa ograničenom ρ funkcijom $\hat{\beta}$ iznosi $\varepsilon^* \leq \varepsilon^*_{max} := \frac{n-p}{2n}$, pa kada $n \to \infty$ dobija se $\varepsilon^*_{max} \approx 0.5$.

U okviru ove klsase ocenjivača najinteresantnije su upravo *MM* i *S* klase ocenjivača i detaljnije će biti diskutovane nešto kasnije.

Vratimo se na analizu rešavanja problema (3.5) i (3.6). Ipak, ukoliko nas ocena skale, S_n, ne iteresuje ili je unapred poznata ocena, onda važe sledeće teoreme.

Teorema 29. *Neka je $\rho(r)$ neprekidna, neopadajuća i neograničena funkcija po $|r|$. Onda postoji rešenje problema* (3.3).

Teorema 30. *Pretpostavimo da je ψ neopadajuća. Onda za dati uzorak (x_i, y_i) obima n*

$$L(\beta) = \sum_{i=1}^{n} \psi(\frac{y_i - x_i^T\hat{\beta}}{S_n})x_i.$$

Tada važi sledeće

- *Sva rešenja $L(\beta) = \mathbf{0}$ su i rešenja problema* (3.3).
- *Ako dalje $\frac{\partial}{\partial\beta}\psi > 0$, onda $L(\beta) = \mathbf{0}$ ima jedinstveno rešenje.*

Teorema pokazuje jedinstvenost monotonih *M* ocenjivača.

3.4 *GM* ocenjivači

Budući da klasični *M* ocenjivači nisu (kvalitativno) robusni (kasnije će biti pokazana neograničenost uticajne funkcije) u odnosu na tačke jakog uticaja dolazi do razvoja klase generalizovanih *M* ocenjivača (nadalje GM) definisanih sa

$$\sum_{i=1}^{n} \eta(x_i, \frac{r_i}{S_n})x_i = \mathbf{0}, \tag{3.9}$$

$$\sum_{i=1}^{n} \chi(\frac{r_i(\hat{\beta})}{S_n}) = 0. \tag{3.10}$$

Funkcija η je najčešće data u obiliku $\eta(x, r) = \pi(x)\psi(r)$ (Mallows, 1975) ili $\eta(x, r) = \pi(x)\psi(\frac{r}{\pi(x)})$ gde funkcijom $\pi(x)$ smanjujemo uticaj tačaka leveridža.

Na slici 3.1 ilustrovana je ta ideja. Marona i Yohai [30] pokazuju postojanje rešenja, jedistvenosti rešenja i asimptotsku nezavisnost (T_n zavisi samo od η, dok S_n zavisi samo od χ) i normalnosti statistika T_n i S_n pod nekim opštim uslovima regularnosti. Detaljnije o ovoj klasi ocenjivača, uticajnoj funkciji, asimptotskim osobinama moguće je naći u [6] i [22].

Pokazano je da za glatku funkciju funkciju η iz (3.9) koja zadovoljava $\eta(x,0)=0$ postoji težinska funkcija w koja definiše (3.8) i daje ista rešenja problema. Ova ekvivalencija je data sa

$$\eta(x,r)=w(x,r)r.$$

Pretpostavimo da $[x^T,y]$ ima raspodelu G na R^{p+1} za koju postoje očekivanja

$$\gamma(G)=E_G[yx], \tag{3.11}$$

i

$$\Sigma(G)=E_G[xx^T]. \tag{3.12}$$

Definicija 76. Pod pretpostavkom da je $\Sigma(G)$ invertibilna, onda je funkcionela za ocenu parametra β metode najmanjih kvadrata definisana sa

$$T(G)=\Sigma^{-1}(G)\gamma(G).$$

Dovoljan uslov za invertibilnost $\Sigma(G)$ jeste da slučajni deo promenljive $x=[1,\ x_1,\ ...,\ x_p]^T$ ima nesingularanu raspodelu. Singularne raspodele su raspodele koncentrisane na Lebegovom skupu mere nula, odnosno gde je verovatnoća realizacije svake tačke mere nula (one nemaju funkciju gustine). Singularne raspodele nisu apsolutno neprekidne u odnosu na Lebegovu meru.

Primer 17. *Za konkertan uzorak obima n dat sa $[x_i^T,y_i]$ matrice $[X|Y]$ empirijska funkcija raspodele F_n koja stavlja masu $\frac{1}{n}$ na svaku observaciju $[x_i^T,y_i]$ daje*

$$\gamma(F_n)=\frac{1}{n}\sum_{i=1}^{n}y_ix_i=\frac{1}{n}X^TY$$

i

$$\Sigma(F_n)=\frac{1}{n}\sum_{i=1}^{n}x_ix_i^T=\frac{1}{n}X^TX,$$

odnosno,

$$T(F_n)=\Sigma^{-1}(F_n)\gamma(F_n)=(X^TX)^{-1}X^TY,$$

što je upravo izraz za $\hat{\beta}$ metode LS.

3.5 Uticajna funkcija metode *LS*

Po definiciji uticajna funkcija je izvod po pravcu $[x^Ty]$ funkcionele T u raspodeli F. Neka je $H=\delta_{[x^Ty]}-F$, razlika u dve raspodele. Tada je

$$T(F+tH)=\Sigma_{F+tH}^{-1}(\gamma_{F+tH})=(\Sigma_F+\Sigma_{tH})^{-1}(\gamma_F+t\gamma_H).$$

Odatle sledi

$$\frac{\partial}{\partial t}[T(F+tH)]_{t=0}$$
$$= \Big[[\frac{\partial}{\partial t}(\Sigma_F + t\Sigma_H)^{-1}](\gamma_F + t\gamma_H) + (\Sigma_F + \Sigma_{tH})^{-1}\gamma_H\Big]_{t=0}$$
$$= \Sigma_F^{-1}\gamma_H - \Sigma_F^{-1}\Sigma_H\Sigma_F^{-1}\gamma_F = \Sigma_F^{-1}(\gamma_H - \Sigma_H\Sigma_F^{-1}\gamma_F)$$
$$= \Sigma_F^{-1}(\gamma_H - \Sigma_H T(F)),$$

jer je

$$\frac{\partial}{\partial t}[(\Sigma_F + t\Sigma_H)^{-1}] = -\Sigma_F^{-1}\Sigma_H\Sigma_F^{-1}.$$

Na kraju se dobija

$$\frac{\partial}{\partial t}[T(F+tH)]_{t=0} = \Sigma_F^{-1}E_H[xy - xx^T(F)] = E_H[IF(x^T, y, F)]$$

gde je

$$IF(x^T, y, F) = \Sigma_F^{-1}[x(y - x^T T(F))]. \tag{3.13}$$

Primetimo da u (3.13) figuriše uticaj observacija preko $\Sigma_F^{-1}x$ i uticaj reziduala preko $(y - x^T T(F))$.

Tejlorovim razvojem funkcionele $\hat{\beta}$ u okolini β dobija se

$$\hat{\beta} = T(F_n) = T(F) + \frac{1}{n}\sum_{i=1}^{n} IF(x_i^T, y_i, F) + R_n,$$

odnosno

$$\sqrt{n}[T(F_n) - T(F)] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF(x_i^T, y_i, F) + \sqrt{n}R_n,$$

gde je R_n ostatak Tejlorovog razvoja. Ako $\left\|\sqrt{n}R_n\right\| \xrightarrow{p} 0$ onda iz centralne granične teoreme sledi

$$\sqrt{n}[T(F_n) - T(F)] \xrightarrow{d} N(0, Cov[F, F]),$$

gde je

$$Cov[F, F] = E_F[IFIF^T] = \Sigma_F^{-1}E_F[\epsilon^2] = \sigma^2\Sigma_F^{-1}.$$

Dakle, problem asimptotske normalnosti je sveden na problem $\left\|\sqrt{n}R_n\right\| \xrightarrow{p} 0$.

Teorema 31. *Neka je F_n empirijska distiribucija tačaka $[x_i^T, y_i]$, $i = 1, ..., n$ koje su* i.i.d. *sa funkcijom raspodele F_x za svaki p dimenzionalni vektor x. Pretpostavimo linearnu vezu $y = x^T\beta - \epsilon$ i nezavisnost slučajne promenljive x i ϵ i F_ϵ je simetrična sa centrom simetrije u 0. Pretpostavimo dalje da matrica $\Sigma(F) = E_F[xx^T]$ postoji i invertibilna je i da postoji $\gamma(F)$. Ukoliko važi*

$$n^{1/4}[\gamma(F_n) - \gamma(F)] = O_p(1),$$

i važi

$$n^{1/4}[\Sigma(F_n) - \Sigma(F)] = o_p(1).$$

Pod ovim uslovima važi $\sqrt{n}\|R_n\| \xrightarrow{p} 0$, kada $n \to \infty$, odnosno

$$\sqrt{n}[T(F_n) - T(F)] \xrightarrow{d} N(0, \sigma^2\Sigma_F^{-1}).$$

3.5.1 Kriterijumi za poređenje višeparametarskih ocenjivača

Navodimo tri kriterijuma za poređenje ovih ocenjivača na konačnim uzorcima. Pretpostavimo da su nam data dva nepristrasna ocenjivača parametra β, $\hat{\beta}_1$ sa $\Sigma_1 = Cov[\hat{\beta}_1, \hat{\beta}_1^T]$ i $\hat{\beta}_2$ sa $\Sigma_2 = Cov[\hat{\beta}_2, \hat{\beta}_2^T]$.

Ocenjivač $\hat{\beta}_1$ je efikasniji ukoliko važi bar jedan od sledećih uslova:

- $tr(\Sigma_1) < tr(\Sigma_2)$.
- $\Sigma_2 - \Sigma_1$ je pozitivno semidefinitna matrica.
- $det(\Sigma_1) < det(\Sigma_2)$.

Pretpostavimo da postoji $\beta \in R^p$, odnosno da komponente $x^T,\ y$ iz $[x^T, y]$ zadovoljavaju sledeće:

(i) $E_G[y|x] = x^T\beta$

(ii) $\epsilon = y - x^T\beta$ je nezavisno u odnosu na $x^T\beta$

(iii) G_ϵ je simetrična oko 0 sa konačnom varijasnom σ^2.

Ove pretpostavke se zovu regresiona struktura.

Lema 11. *Ako je $T(G)$ dobro definisana ocena parametra β onda je ona Fišer konzistentna.*

Podsetimo se još jednom definicije rezidual *LS* metoda

$$r(x^T, y, G) = y - x^T T(G). \tag{3.14}$$

Interesuju nas funkcije težinske funkcije $w = w(x, r)$ gde je r dato sa 3.14. Napominjemo da težina dodeljena tački $[x^T, y]$ ne zavisi samo od tačke, već i od raspodele G jer je položaj tačke relativan u odnosu na G i rezidual u tački zavisi od ocene reziduala $T(G)$.

3.6 Asimptotska teorija težinskih najmanjih kvadrata

Analogno slučaju metode klasičnih najmanjih kvadrata definišu se očekivanja

$$\gamma_w(G) = E_G[wyx] \tag{3.15}$$

$$\Sigma_w(G) = E_G[wxx^T] \tag{3.16}$$

odakle sledi funkcionela za ocenu β metodom težinskih kvadrata,

$$T_w(G) = \Sigma_w^{-1}(G)\gamma_w(G).$$

Za pomenutu empirijsku raspodelu F_n γ_w je data kao

$$\gamma_w(F_n) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i - x^T\beta)y_i x_i,$$

gde su nepoznati parametri β i može se definisati statistika

$$\gamma_{\hat{w}}(F_n) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i - x^T T(F_n)) y_i x_i.$$

Indeks w označava poznati parametar, dok $\hat{w}$ označava ocenjeni parametar. Analogno se razlikuju $\Sigma_w(F_n)$ i $\Sigma_{\hat{w}}(F_n)$ i $T_w(F_n)$ i $T_{\hat{w}}(F_n)$. Elegantnije zapisano ove statistike date su sa

$$\gamma_{\hat{w}} = \frac{1}{n} X^T \hat{W} Y.$$

$$\Sigma_{\hat{w}}(F_n) = \frac{1}{n}(X^T \hat{W} X)$$

$$T_{\hat{w}}(F_n) = (X^T \hat{W} X)^{-1} X^T \hat{W} Y$$

gde je $\hat{W}$ dijagonalna matrica sa $\hat{W}_{ii} = w(x_i, y_i - x^T T(F_n))$.

Teorema 32 (*wLSE*)**.**

(i) Pretpostavimo da su zadovoljeni uslovi za postojanje $\gamma(G)$, $\Sigma(G)$ i $T(G)$ u 3.11 *i* (3.12)*. Neka je r definisano sa* (3.14) *i pretpostavimo da je $w = w(x, r)$ je nenegativna, ograničena i merljiva funkcija po (x, r). Onda su $\gamma_w(G) = E_G[wyx]$, $\Sigma_w(G) = E_G[wxx^T]$ dobro definisane.*

(ii) Pretpostavimo iz (i) važi $w \geq 0$. Ako za svaki nenula vektor a skup

$$\{[x^T, y] \ : a^T x \neq 0 \text{ i } w(x, y - x^T\beta) > 0\}$$

ime pozitivnu verovatnoću za G, onda egzistencija $T(G)$ implicira egzistenciju $T_w(G)$.

(iii) Ako dodatno uz pretpostavke (i) i (ii), $G = F_\beta$ ima strukturu regresije i važi simetričnost w tj $w(x, r) = w(x, -r)$ za svako x, r, onda postoji $T_w(F_\beta) = \beta$, odnosno T_w je Fišer konzistentno za β.

(iv) Ako dodatno uz uz pretpostavke (i) i (ii), pretpostavimo da za komponente x postoji četvrti momenat i da težinska funkcija mora da zadovoljava

$$\big|w(x, q) - w(x, r)\big| \leq c|q - r|$$

za neku pozitivnu konstantu c i svako r i q, onda su wLSE *ocenjivači slabo konzistentni odnosno*

$$\hat{\beta}_{\hat{w},n} \xrightarrow{p} \beta.$$

3.7 Uticajna funkcija metode *wLSE*

Funkcionela težinskih najmanjih kvadrata T_w je definisana u tri koraka. Prvi korak, pođemo od Fišer konzistentnog ocenjivača $T(G)$. U drugom koraku računamo reziduale u tački $[x, y]$ raspodele G iz (3.14) i u trećem koraku računamo ocenjivač metode *wLSE* $T_w(G)$.

Uticajna funkcija funkcionele $T_w(G)$ u F u smeru tačaka $\delta_{[x^T,y]}$ je data izrazom

$$IF_{T_w}(x^T, y, F) = \Sigma_w^{-1}(wrx) - \Sigma_w^{-1} C_w IF_T(x^T, y, F)$$

gde je $C_w = E_F[\frac{\partial}{\partial r} w(x, r) r x x^T]$. Detaljno izvođenje ove formule može se naći u [22] i [32] na 255 str.

Primetimo da je razlika u uticajnoj funkciji ovog metoda u odnosu na metodu *LS*razlikuje u dve komponente. Prva je $\Sigma_w^{-1}(wrx)$ gde je Σ_w^{-1} i $w = w(x, r)$ koji smanjuju uticaj tačaka sa visokim leveridžom i time omogućavaju uticajnoj funkciji da ostane ograničena. Druga komponenta se ispoljava u matrici $\Sigma_w^{-1} C_w$. Ako je težinska funkcija $w(x, r)$ odabrana tako da je norma matrice $\Sigma_w^{-1} C_w$ mala, onda se uticaj $IF_T(x^T, y, F)$ može smanjiti, ali ne i u potpunosti eliminisati. Druga komponenta je ograničena ako je ili $C_w = 0$ ili inicijalni ocenjivač ima ograničenu uticajnu funkciju. Ukoliko je $IF_T(x^T, y, F)$ neograničena, onda će i $IF_{T_w}(x^T, y, F)$ biti neograničena.

Takođe u [22] se može naći detaljnije o uslovima pod kojim uslovima je T_w dobro definisan i konzistentan ocenjivač. Analizirana je konvergencija *GM* ocenjivača T_w i izražena uticajna funkcija u $k + 1$-oj iteraciji. Ova veza dalje omogućava analizu uslova pod kojima iterativna metoda $T_k(G)$ konverigra ka rešenju bez obzira na početni ocenjivač.

Analizirana je i implementacija *M* ocenjivača sa Njutnov-Rapsonovom metodom i donet je zaključak da se asimptotski ponaša kao i ocenjivač koji je implementiran iterativno preko IRWLS metode.

3.8 *S* ocenjivači

S ocenjivači formiraju klasu ocenjivača regresije koja poseduje visoku tačku preloma. Ova klasa ocenjivača koristi ograničene ρ funkcije i afino invarijantna, skalarno i regresiono ekvivalentna. Iz ograničenosti sledi da je ψ reopadajuća. Na isti način na koji klasa *LS* minimizira varijanse reziduala, klasa *S* ocenjivača minimizira disperziju reziduala, $s(r_1(\beta), ..., r_n(\beta))$.

Definicija 77. Disperzija, s, reziduala data je kao rešenje problema

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\frac{r_i}{s}) = K,$$

gde je K konstanta i funkcija ρ zadovoljava sledeće uslove

1. $\rho(0) = 0$.
2. ρ je simetrična i neprekidno diferencijabilna.

3. Postoji konstanta $a > 0$ tako da je ρ strogo rastuća na $[0, a]$ i konstantna na $[a, \infty)$.
4. $\frac{K}{\rho(a)} = \frac{1}{2}$.

Četvrti uslov nije nužan, ali upravo on obezbeđuje tačku preloma od 0.5. Konstanta K se često bira tako da rezultujuće s je ocena σ kada vektor grešaka ima normalnu raspodelu. Dakle, $K = E_\Phi(\rho(r))$. Najčešće se implementiraju uz Takijevu bikvadratnu ρ funkciju.

3.9 *MM* ocenjivači

Klasa *MM* ocenjivača kombinuje klasu *M* ocenjivača sa klasom *S* ocenjivača i kao produkt dobija se klasa ocenjivača koja istovremeno ima visoku tačku preloma i poseduje visoku efikasnost. Neka je ρ realna funkcija koja zadovoljava sledeće pretpostavke.

(A) (i) $\rho(0) = 0$ i $\rho(\frac{0}{0}) = 0$ po definiciji;
(ii) $\rho(-r) = \rho(r)$;
(iii) $0 \le r \le v \;\Rightarrow\; \rho(r) \le \rho(v)$;
(vi) $a = \sup \rho$, onda $0 < a < \infty$;
(v) ako $\rho(u) < a$ i $0 \le v$, onda $\rho(u) < \rho(v)$.

Za dati uzorak obima n, ocenivač skale S_n definisan je kao rešenje

$$\frac{1}{n}\sum_{i=1}^{n} \rho(\frac{r_i}{S_n}) = b \tag{3.17}$$

gde je b definisano preko $E_\Phi[\rho(r)] = b$, a Φ je standardna normalna raspodela.

Ukoliko

$$c(u) = \frac{card\{i : 1 \le i \le n,\ r_i = 0\}}{n} < 1 - \frac{b}{a},$$

onda problem (3.17) ima jedinstveno rešenje različito od nule. Inače, ukoliko je $c(u) \ge 1 - \frac{b}{a}$, onda definišemo $s(u) = 0$.

MM ocenjivači su definisani u tri koraka.

Korak 1. Neka je $\hat{\beta}_{0,n}$ inicijalni ocenjivač sa što većom tačkom preloma, ukoliko je moguće želimo da tačka premoma iznosi $\varepsilon^* = 0.5$. Često se za početni ocenjivač uzima upravo klasa *S* ocenjivača.

Korak 2. Prvo se izračunaju reziduali

$$r(\hat{\beta}_{0,n}) = y - \hat{\beta}_{0,n} x_i.$$

Potom se izračunava *M*-ocenjivač skale $S_n = S_n(r(\hat{\beta}_{0,n}))$ definisan sa (3.17) funkcijom ρ_0 koja zadovoljava navedene pretpostavke uz konstante a i b tako da važi $\frac{b}{a} = 0.5$, gde je $a = \max \rho_0(r)$. Ova pretpostavka implicira da je tačka preloma ocenjivača skale S_n iznosi $\varepsilon^* = 0.5$.

Korak 3. Neka je ρ_1 druga funkcija koja zadovoljava pretpostavke (A) tako da

$$\rho_1(r) \leq \rho_0(r), \tag{3.18}$$

$$\sup \rho_1(r) = \rho_0(r) = a. \tag{3.19}$$

Neka je $\psi_1 = \frac{\partial}{\partial r}\rho_1$. Onda su *MM* ocenjivači $\hat{\beta}_{1,n}$ definisani kao rešenje problema

$$\sum_{i=1}^{n} \psi_1(\frac{r_i}{S_n})x_i = \mathbf{0}, \tag{3.20}$$

što potvrđuje

$$S(\hat{\beta}_{1,n}) \leq S(\hat{\beta}_{0,n}),$$

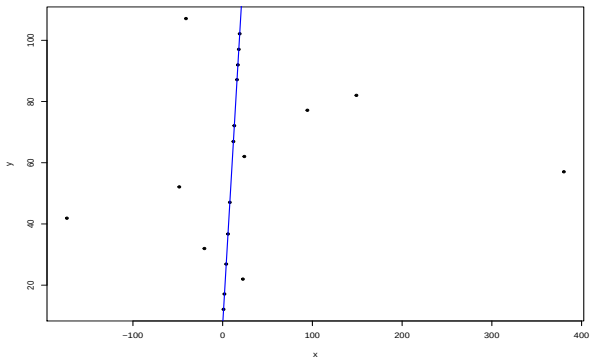
gde je

$$S(\beta) = \sum_{i=1}^{n} \rho(\frac{r_i(\beta)}{S_n}).$$

Dakle, prva dva koraka su odgovorna za visoku tačku preloma ocenjivača, dok je treći korak odgovoran za visoku efikasnost ocenjivača. Upravo zbog ovoga ρ_0 i ρ_1 ne moraju nužno biti iste funkcije i zašto ocenjivač izabran u ovom koraku ne mora biti nužno efikasan.

Yohai dokazuje da ukoliko se u prvom koraku koristi ocenjivač sa tačkom preloma 0.5 onda će *MM* ocenjivač takođe imati tačku preloma 0.5. Klasa *MM* ocenjivača poseduje *egzakt fit* osobinu[4] koja kaže da na datom uzorku obimu n ukoliko bar $n - \frac{n}{2} + 1$ observacija zadovoljavaju $y = x_i^T\beta$, onda se kao ocenjivač parametra dobija upravo β nezavisno od drugih observacija. Ovu osobinu poseduju robusne klase ocenjivača poput *S*. Takođe ukoliko se u prvom koraku uzme regresiono invarijantan i/ili afino invarijantan ocenjivač onda će i dobijeni *MM* ocenjivač zadržati te osobine.

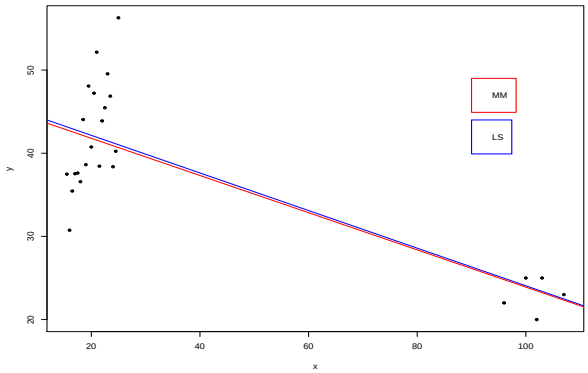
Detaljnije o ovim ocenjivačima u [38]. Takođe tu se nalazi strategija za izbor ρ_0 i ρ_1, asimptotske osobine (konzistentnost i asimptotska normalnost), uticajna kriva, varijansa kao i algoritam za komputaciju preko IRWLS algoritma. Napominjemo da bez obzira na osobine *MM* ocenjivača uticajna funkcija nije ograničena.



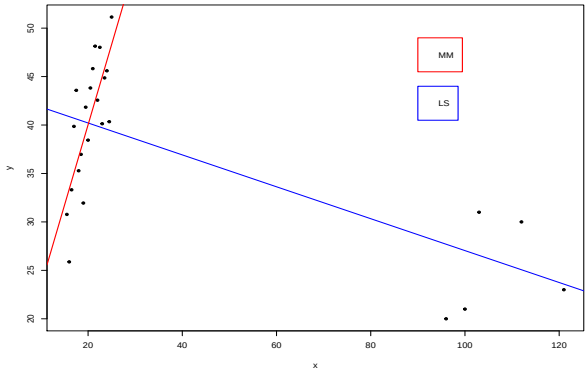
Slika 3.2: Uprkos 9 (od ukupno 20) observacija koje su autlajeri, *MM* ocenjivač pronalazi linearan fit.

[4]Exact fit property.

Bez obzira na ove impresivne osobine *MM* ocenjivača i ovaj ocenjivač se relativno lako pokvari kada su nivoi kontaminacije mnogo manji od 50%. Ovo ilustrujemo primerom.



Slika 3.3: Uprkos kontaminaciji uzorka od svega 20% *MM* ocenjivač se u potpunosti kvari i ponaša se isto kao i ocenjivač metode *LS*.



Slika 3.4: Mala promena uzorka i *MM* ocenjivač uspeva da pronađe liearan trend u podacima.

3.9.1 Izvori problema *MM* ocenjivača

Bez obzira na sposobnost *MM* ocenjivača da izađu na kraj sa individualnim autlajerima u primeru ilustrovanom na Slici 3.4, videli smo osetljivost ocenjivača na klaster podataka u primeru na Slici 3.3.

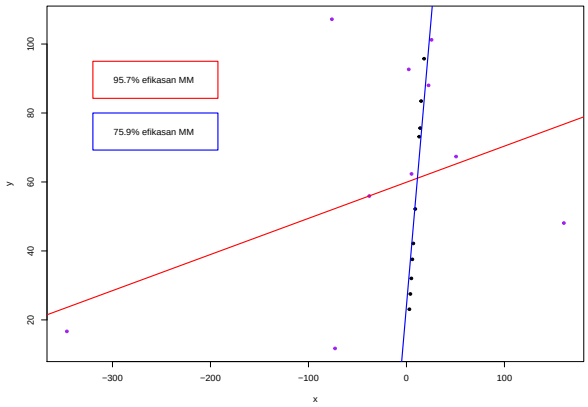
Rousenev i Liroj[5] (1987, str. 154) naglašavaju da je visoka tačka preloma potreban, ali ne i dovoljan uslov za dobar robusan ocenjivač. Deo problema leži u

[5]Rousseeuw i Leroy.

činjenici da tačka preloma samo razmatra uticaj autlajera koji teže u beskonačnost.

Bianko[6] ([2]) objašnjava, ako ocejnivač ima tačku preloma manju od ε, onda ocenjivač ostaje u zatvorenom ograničenom skupu, ali ako deo autlajera prilazi ε ovaj skup postaje jako veliki. Ovo znači da bez obzira što se ocenjivač ne kvari, on može postati vrlo nepouzdan i da bude vrlo udaljen od prave ocene parametra. Robusnost ocenjivača na autlajere donekle zavisi od stope kojom se ovaj skup širi kako se udeo autlajera pocećava.

Drugi izvor loše preformanse *MM* klase ocenjivača može da leži i u Koraku 3. definicije ovog ocenjivača. Upravo izbor konstanti koji obezbeđuje efikasnost i do 95% može da ima takvu posledicu da se ovaj ocenjivač ponaša znatno lošije, odnosno proizvodi lošiji fit.



Slika 3.5: Manje efikasni *MM* ocenjivači proizvode znatno bolji linearan fit.

Metodologija iza oceenjivača skale i selekcije parametara za postizanje efikasnosti ocenjivača je bazirana na asimptotskim rezultatima, odnosno ovi rezultati su bazirani na činjenici da je p fiksirano, a da $n \to \infty$. Ovo znači da na preformans ocenjivača znatno utiče i obim uzorka. Na konančnim populacijama $\frac{p}{n}$ može biti preveliko. Ispostavlja se da i na malim uzorcima gde je količnik $\frac{p}{n}$ dovoljno mali rezultati i dalje važe, inače se javljaju problemi. Maronna i Yohai (2010) pokazuju da ukoliko je količnik $\frac{p}{n}$ dovoljno veliki, onda je efikasnost *MM* ocenjivača lošija iz dva razloga

1. ocenjivač skale greške potcenjuje pristrasnosti, što imlicira da će potceniti i tačnu ocenu skale.

2. stvarna efikasnost je znatno niža od željene, jer parametri ψ funkcija odabrani u skladu sa asimptotskom teorijom su neodgovarajući.

3.10 Numerička implementacija

U ovom poglavlju ilustrujemo numeričke sposobnosti ovih ocenjivača uz neke ograničavajuće faktore. Budući da većina primera u ovom poglavlju dolazi realnog

[6] Bianco.

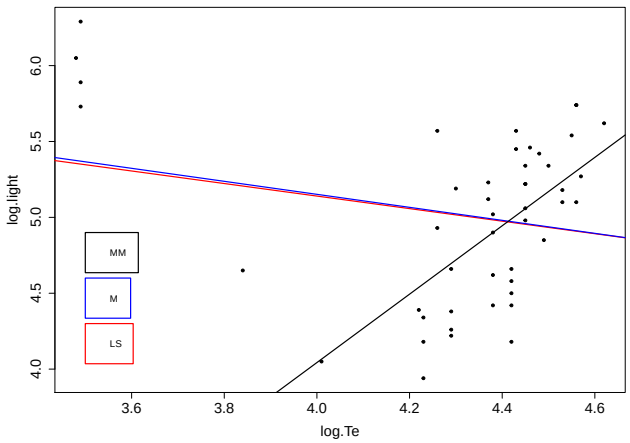
sveta radimo na konačnim uzorcima. Odnosno, u ovom slučaju je količnik $\frac{p}{n}$ mnogo veći nego u asimptotoskim slučevima. Takođe, budući da tačni modeli nisu poznati i budući da nije poznata prava raspodela, kao ni nivoi kontaminacije podataka, ostajemo uskraćeni za preciznu analizu tačnosti ovih ocenjivača. Napominjemo da se na simuliranim podacima bolje i preciznije ispoljavaju osobine metoda kojim se podaci analiziraju. Ipak, uprkos ovome dobićemo uvid u neke pozitivne i negativne strane ovih ocenjivača.

U ovom izveštaju korišćen je paket *MASS* gde su implementirani skalarno ekvivalentni ocenjivači sa nekim njihovim uobičajenim metodama. Takođe, kao ocenjivač skale korišćen je *MADN*. Ukoliko se ne specificiraju željene funkcije *M* ocenjivači koriste Huberovu ψ funkciju sa parametrom $k = 1.345$ i skalu ocenjuju *MAD* ocenjivačem, dok *MM* ocenjivači koriste *S* ocenjivač sa Takijevom bikvadratnom funkcijom sa $a = 1.548$ u prva dva koraka.

Primer 18. *Neka su dati podatci iz Hertzsprung-Russell diagrama klastera zvezda CYG OB1* [7] *koji sadrži 47 zvezda u pravcu sazvežđa Cygnus. Prva promenljiva je logaritam efektivne temperature na površini zvezde X, a druga promenljiva je intezitet svetla Y. Podaci sadrže četiri džinovske zvezde - autlajere koje su uticajne tačke, ali ne i greške.*

Tabela 3.1: Rezultati za podatke Hertzsprung-Russellovog dijagrama.

Metoda	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	6.7935	-0.4133
M	6.8658	-0.4285
MM	-4.9702	2.2533



Slika 3.6: *LS* (crvena), *M* (plava) i *MM* (crna) regresija za podatke Hertzsprung-Russellovog dijagrama klastera zvezda CYG OB1.

[7] Podatke je moguće naći u paketu robustbase pod nazivom *starsCYG*.

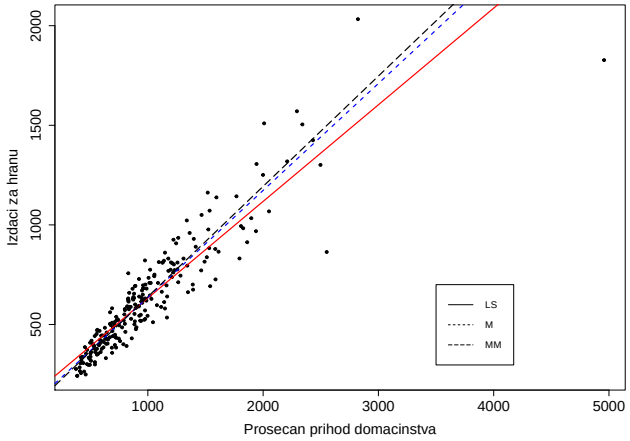
Primetimo da su LS i M ocenjivači podlegli uticaju tačakama jakog leveridža, dok su MM ocenjivači uspeli da nađe odgovarajući linearan trend.

Primer 19. *Koenker i Bassett (1982) daju skup podataka Engel o prosečnim prihodima i izdacima za ishranu*[8]*. U skupu imam 235 registrovane observacije.*

Tabela 3.2: Rezultati za podatke Engel o izdacima za ishranu i prosečnim prihodima.

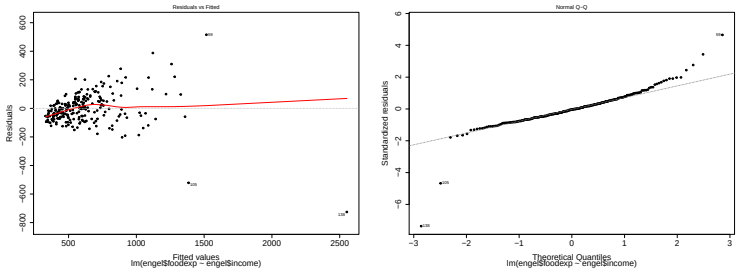
Metoda	$\hat{\beta}_0$	$\hat{\beta}_1$
LS	147.47539	0.48518
M	99.4319	0.5368
MM	85.5072	0.5539
S	104.9857	0.5296

Takođe je moguće izračunavanje S *ocenjivača postavljajući maksimalan broj iteracija za konvergenciju na 0 (*maxit $= 0$*).*

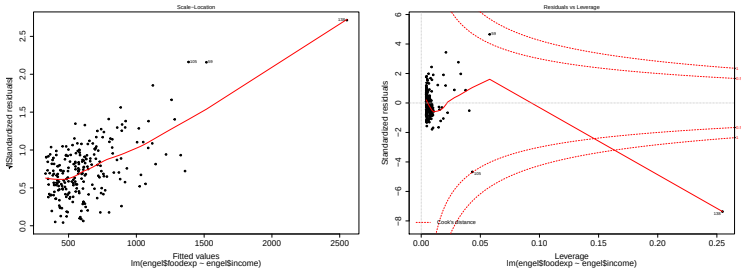


Slika 3.7: *LS* (crvena), *M* (plava) i *MM* (crna) regresija za podatke Engel o izdacima za ishran i prosečnim prihodima domaćinstva.

Primetimo da su LS *ocenjivači podlegli uticaju par autlajera koje se mogu videti sa priloženog* QQ*-plota, dok* MM *ocenjivači nemaju ove probleme i generalno prate trend većine podataka.*



[8]Ovi podaci mogu se naći u paketu quantreg.



Primer 20. *Neka je dat skup podataka* airquality *iz paketa* MASS. *Ovaj skup podataka sadrži merenja o kvalitetu vazduha u Njujorku za 153 dana, počev od 1. Maja do 30. Septembra 1973. Varijable su*

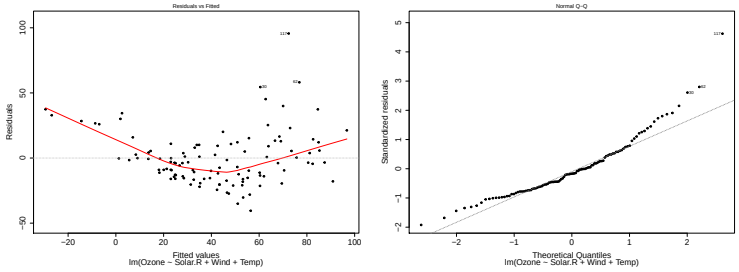
- *Ozon (Ozone) - srednji ozon u* ppb *(koncentracija).*
- *Temperatura (Temp) u stepenima Farenhajta.*
- *Solarna radijacija (Solar.R) u Lenglejs frekventnom opsegu 4000-7700.*
- *Prosečna brzina vetra (Wind) u miljama po satu.*

Posle uklanjanja nekompletnih vrsta observacija u skupu je ostalo 111 observacija. Dobijeni su sledeći rezultati.

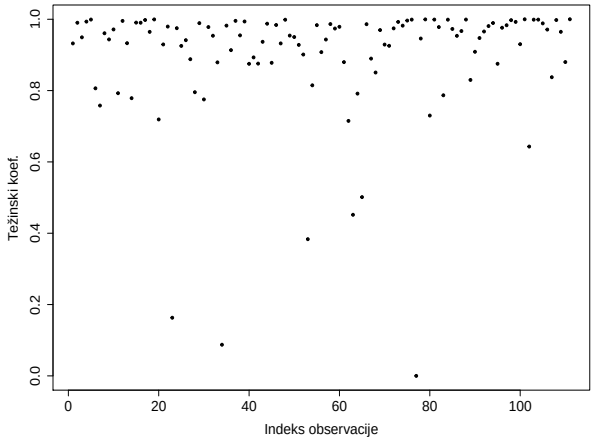
Tabela 3.3: Rezultati za podatke airquality.

Metoda	$\hat{\beta}_I$	$\hat{\beta}_S$	$\hat{\beta}_W$	$\hat{\beta}_T$
LS	-64.34208	0.05982	-3.33359	1.65209
M	-78.4539	0.0493	-2.6438	1.7448
MM	-85.1726	0.0448	-2.2581	1.7820
S	-84.8094	0.0455	-2.2999	1.7794

Primetimo da postoje znatne razlike između $\hat{\beta}_{LS}$ *i* $\hat{\beta}_M$ *i* $\hat{\beta}_{MM}$. *Što implicira da je ocenjivač* LS *najverovatnije podlegao uticaju tačaka jakog leveridža ili autlajera. Isto se može konstatovati i za* M *ocenjivač.*



Slika 3.8: Residuali vs fitovane vr. plot i QQ-normal plot.



Slika 3.9: Težinski koeficijenti observacija.

Observacije sa težinskim koeficijentima manjim od 0.6 su observacije pod rednim brojevima 23, 34, 53, 63, 65, 77. Ove observacije je potrebno dalje ispitati.

Primer 21. *Neka su dati podaci o 23 aviona sa jednim motorom*[9]*. Promenljive koje figurišu u modelu koji pravimo su*

- X_1 *zavisna varijabla koja predstavlja cenu ovih aviona (u* $100.000\$$*).*
- X_2 *Lift to drag ratio.*
- X_3 *je težina aviona.*

Razliku u obimu odgovarajućih MM *i* LS *ocenjivača merimo*

$$\left\|\hat{\beta}_{MM} - \hat{\beta}_{LS}\right\|_2.$$

Tabela 3.4: Rezultati.

Metoda	$\hat{\beta}_I$	$\hat{\beta}_{LTD}$	$\hat{\beta}_W$	$\left\|\hat{\beta}_{MM} - \hat{\beta}_{LS}\right\|_2$
LS	5.512	0.054	-0.0000970	-
95% efikasan *MM*	5.495	0.041	-0.0000948	0.021
91.7% efikasan *MM*	5.486	0.038	-0.0000941	0.031
84.7% efikasan *MM*	5.131	1.653	-0.0002932	1.644
75.9% efikasan *MM*	5.036	1.833	-0.0003131	1.842
66.1% efikasan *MM*	4.935	1.981	-0.0003285	2.012

Zapazimo sličnu preformansu 95% i 91.7% efikasnih MM *ocenjivača sa ocenjivačima* LS*. Primetimo da kada se asimptotska efikasnost ocenjivača smanji, ponašanje* MM *ocenjivača se znatno razlikuje od metode* LS*.*

[9]Podaci su dostupni u paketu robustbase pod nazivom airplane.

Glava 4

Teorija kredibiliteta

Osnovna ideja iza osiguranja bavi se upravo minimizacijom određenog stresa u smislu minimizacije efekata nepredviđenih događaja koji se odražavaju na konkretnog pojedinica ili grupu pojedianca. Odnosno, naše društvo učinjeno je robusnim u odnosu na određenu vrstu rizika. Ovaj cilj postignut je decentralizacijom, odnosno diversifikacijom rizika. Rizik definišemo kao neizvesnost, koja proizilazi iz događaja, čiji efekat je moguće finansijski kvantifikovati. Rizik je sa pojedinca, nosioca polise, prebačen na profitom motivisano osiguravajuće društvo.

4.1 Analiza kredibiliteta

Razvojem osiguranja kao industrije i nauke nastajale su i razvijale se različite naučne metode za procenu rizika. Teorija kredibiliteta je grana aktuarske nauke koja se bavi kvantifikacijom potencijalnih budućih rizika na osnovu danas dostupnih podataka osiguranika. Preciznije, reč je o tehnici koja određuje premiju konkretnog nosioca ugovora tako što postiže balans između očekivanja potraživanja konkretnog ugovora sa očekivanim potraživanjem portfolia sličnih ugovora. Cilj je što preciznije odrediti premiju konkretnog nosioca ugovora.

Postoji više pristupa ovoj problematici, stoga pominjemo neke istorijski značajne modele. Prvi iskorak pravi Mowbary (1914) koji uvodi model ograničene fluktuacije kredibilnosti. Ovaj model ima za cilj da odredi potrebno izlaganje riziku (u smislu hronološki dostupnih informacija o kredibilnosti nosioca ugovora). Nosiocima polise koji nemaju pun kredibilitet dodeljuje se parcijalni kredibilitet, a ostatak premije se obračunava pod uticajem kredibilnosti portfolia osiguranika. Premija se dobija kao konveksna kombinacija premije osiguranika, $\hat{X}$, i manuelne premije (premije portfolija), M, odnosno

$$P_c = Z\hat{X} + (1 - Z)M.$$

Ovaj model nije precizan i ima dosta mana, ali predstavlja bitan iskorak, detaljnije o modelu može se pročitati u [37].

Drugi pristup kredibilnosti je preko Bajezove statistike. Ovde je reč o metodima najveće tačnosti i najbitnije je pomenuti Bulmanov i Bulman-Straubov model. Ukratko u zavisnosti od karakteristike rizika, rizici su grupisani na razne riziko klase

i premija mora biti izračunata za svaku od klasa. Dakle, pretpostavlja se da postoji slučajna promeljiva Θ koja modelira rizik, čija je realizacija upravo riziko klasa $\Theta_j, j = 1, ..., K$. Detaljnije o ovim modelima može se pročitati u [37].

Predmet rada se bavi multidimenzionalnim kredibilitetom koji je implementiran preko regresije. Najviše pažnje biće posvećeno Hachemeisterovom modelu koji je ekstenzija multidimenzionog kredibilteta.

Međutim procena kredibilnosti ostaje osetljiva na autlajere u podacima, stoga upravo motivacija za ovaj rad proizilazi iz nedostataka robusnih ocenjivača kredibiliteta. Što je grupa podataka manja to je teže detektovati autlajere kao i zakon raspodele iz kog podaci dolaze. U ovoj glavi prezentovan je model regresionih ocenjivača kredibiliteta i potom njegova robusna verzija.

4.2 Model kredibiliteta preko regresije

Neka je dat portfolio od K rizika i neka je

$$\boldsymbol{Y_i}' = (Y_{i1}, ..., Y_{in})$$

vektor observacija i-tog rizika. Napominjemo da neke realizacije rizika Y_{i1} mogu biti date vezom $Y_{i1} = f(Z)$, gde je $f(Z)$ neka funkcija. Neka je $\boldsymbol{w_i}' = (w_{i1}, ..., w_{in})$ vektor poznatih težina asociran sa i-tim rizikom. Na primer Y_{ij} mogu biti prosečna potraživanja i-tog stanje u j-tom kvartalu i w_{ij} su odgovarajući ukupni broj ovih portaživanja kao što je slučaj u Hatchmeisterovom modelu.

Pretpostavimo da dati ugovori $(\Theta_i, \boldsymbol{Y_i})$ zavodovljavaju uslove regresije navedene u Glavi 3. Ova pretpostavka vodi do postojanja vektora β koji nije fiksiran već ga posmatramo kao slučajnu promenljivu koja ima perturbacije determinisane od strane kolektiva, odnosno portfolia. Dakle za datu riziko klasu Θ_i potrebno je oceniti vektor $\beta(\Theta_i)$.

4.2.1 Model standardne regresije

Neka je rizik i okarakterisan od strane riziko klase Θ_i, koja je realizacija slučajne promenljive Θ. Pretpostavimo da važi sledeće:

$R1$ Uslovno, za dato Θ_i, rizici $Y_{ij}, j = 1, 2, ..., n$ su nezavisni i važi

$$E[\boldsymbol{Y_i}|\Theta_i] = X_i\boldsymbol{\beta}(\Theta_i),$$

gde je

$\beta(\Theta_i)$ vektor regresije koji je potrebno oceniti dužine $p \leq n$,

X_i je poznata nesingularna matrica ranga p,

$Var[Y_{ij}|\Theta_i] = \frac{\sigma^2(\Theta_i)}{w_{ij}}$.

$R2$ Ugovori $(\Theta_1, \boldsymbol{Y_1}), (\Theta_2, \boldsymbol{Y_2}), \ldots$ su nezavisne, a riziko klase $(\Theta_1, \Theta_2, \ldots)$ su nezavisne i identično raspodeljene slučajne promenljive.

Napomene:

- Pretpostavka $R1$ znači da za uslovno dato Θ_i observacije Y_{ij} $(j = 1, ..., n)$ zadovoljavaju klasični model težinskih najmanjih kvadrata.
- Iz uslova $R2$ i uslova o nezavisnosti sledi da je

$$Cov[\boldsymbol{Y_i}, \boldsymbol{Y_i}'|\Theta_i] = \sigma^2(\Theta_i)W_i^{-1},$$

gde je dijagonalna matrica data sa $W_i = \begin{bmatrix} w_{i1} & & & \\ & w_{i2} & & \\ & & \ddots & \\ & & & w_{in} \end{bmatrix}$.

Cilj je za neki dati vektor $a = (a_1, ..., a_p)$ odrediti individualnu premiju

$$\mu_a(\Theta_i) = E[X_a|\Theta_i].$$

Za dati model premija je oblika $\mu_a(\Theta_i) = a^T\beta(\Theta_i)$.
Uvedimo sledeće oznake

$$\beta := E[\beta(\Theta_i)],$$
$$\sigma^2 := E[\sigma(\Theta_i)],$$
$$T := [Cov(\beta(\Theta_i), \beta(\Theta_i)'].$$

Teorema 33. *Pod navedenim pretpostavkama $R1$ i $R2$ važi sledeće:*

1. *BLUE ocenjivač parametra $\beta(\Theta_i)$ dat je sa*

$$B_i = (X_i^T W_i X_i)^{-1} X_i W_i \boldsymbol{Y_i}.$$

2. *Kvadratna matrica gubitka data je sa*

$$E[(B_i - \beta(\Theta_i))(B_i - \beta(\Theta_i))^T] = \sigma^2(X_i W_i X_i)^{-1}.$$

Teorema 34 (Ocenjivač kredibilnosti u standardnom slučaju)**.** *Pod pretpostavkama u standardnom regresionom modelu $R1$ i $R2$, sledi da ocenjivač kredibiliteta za $\beta(\Theta_i)$ zadovoljava*

$$\widehat{\widehat{\beta(\Theta_i)}} = A_i B_i + (I - A_i)\beta$$

gde je

$$A_i = T(T + \sigma^2(X_i^T W_i X_i)^{-1})^{-1},$$

a matrica kvadratnog gubitka data je sa

$$E\big[(\widehat{\widehat{\beta(\Theta_i)}} - \beta(\Theta_i))(\widehat{\widehat{\beta(\Theta_i)}} - \beta(\Theta_i))^T\big] = (I - A_i)T.$$

4.2.2 Model uopštene regresije (Hachemeisterov model)

Jedina razlika ovog i prethodnog modela jeste da relaksiramo uslov nezavisnosti između $\boldsymbol{Y_i}$, odnosno sada dozvaljamo bilo kakvu strukturnu zavisnost između registrovanih vrednosti $\boldsymbol{Y_i}$.

Neka je rizik i okarakterisan riziko klasom Θ_i, koja je realizacija slučajne promenljive Θ. Pretpostavimo da važi sledeće:

$H1$ Uslovno za datu riziko klasu Θ_i, rizici $Y_{ij}, j = 1, 2, ..., n$ su nezavisni i važi

$$E[\boldsymbol{Y_i}|\Theta_i] = X_i\boldsymbol{\beta}(\Theta_i),$$

gde je

$\beta(\Theta_i)$ vektor regresije koji je potrebno oceniti dimenzije $p \leq n$,

X_i je poznata nesingularna matrica ranga p,

i važi

$$Cov[\boldsymbol{Y_i}, \boldsymbol{Y_i}^T|\Theta_i] = \hat{W}_i(\Theta_i), \tag{4.1}$$

gde je $\hat{W}_i$ simetrična pozitivno semidefinitna matrica težinskih koeficijenata.

$H2$ Promenljive $(\Theta_i, \boldsymbol{Y_i}), (\Theta_2, \boldsymbol{Y_2}), ...$ su nezavisne, a $(\Theta_1, \Theta_2, ...)$ su nezavisne i identično raspodeljene slučajne promenljive.

Izraz (4.1) implicira da sada imamo više strukturalne parametre

$$S_i = E[\hat{W}_i(\Theta_i)], i = 1, 2, ..., K$$

umesto $\sigma^2 W_i^{-1}$.

Teorema 35. *Pod navedenim pretpostavkama $H1$ i $H2$ važi sledeće:*

1. *BLUE ocenjivač parametra $\beta(\Theta_i)$ dat je sa*

$$B_i = (X_i^T S_i^{-1} X_i)^{-1} X_i S_i^{-1} \boldsymbol{Y_i}.$$

2. *Matrica kvadratnog gubitka data je sa*

$$E[(B_i - \beta(\Theta_i))(B_i - \beta(\Theta_i))^T] = \sigma(X_i^T S_i^{-1} X_i)^{-1}.$$

Ocenjivač B_i je individualno nepristrasan, odnosno

$$\begin{aligned} E[B_i|\Theta_i] &= (X_i^T S_i^{-1} X_i)^{-1} X_i S_i^{-1} E[Y_i|\Theta_i] \\ &= (X_i^T S_i^{-1} X_i)^{-1} X_i S_i^{-1} X_i \beta(\Theta_i) \\ &= \beta(\Theta_i) \end{aligned}$$

Teorema 36 (Hachemeisterova formula)**.** *Pod pretpostavkama iz modela uopštene regresije $R1$ i $R2$ sledi da ocenjivač kredibiliteta nepoznatog parametra $\beta(\Theta_i)$ zadovoljava*

$$\begin{aligned} \widehat{\widehat{\beta(\Theta_i)}} &= A_i B_i + (I - A_i)\beta \\ &= \beta + A_i(B_i - \beta) \end{aligned} \tag{4.2}$$

gde je

$$A_i = T(T + \sigma^2(X_i^T S_i^{-1} X_i)^{-1})^{-1}, \tag{4.3}$$
$$B_i = (X_i^T S_i^{-1} X_i)^{-1} X_i^T S_i^{-1} \boldsymbol{Y_i},$$
$$S_i = E[\hat{W}_i(\Theta_i)] = E[Cov[\boldsymbol{Y_i}, \boldsymbol{Y_i}^T | \Theta_i)],$$
$$T = Cov[\beta(\Theta_i), \beta(\Theta_i)^T],$$
$$\beta = E[\beta(\Theta_i)].$$

Matrica kvadratnog gubitka data je sa

$$E\big[(\widehat{\widehat{\beta(\Theta_i)}} - \beta(\Theta_i))(\widehat{\widehat{\beta(\Theta_i)}} - \beta(\Theta_i))^T\big] = (I - A_i)T.$$

De Vylder [9] daje ocene parametara b

$$\hat{\beta} = (\sum_{j=1}^{I} A_j)^{-1} \sum_{j=1}^{I} A_j Bj,$$

gde je A_i dato izrazom (4.3), a ocena T data je sa

$$\hat{T} = \frac{1}{K-1} \sum_{j=1}^{I} A_j (B_j - \hat{\beta})(B_j - \hat{\beta})^T.$$

4.3 Robusni regresioni kredibilitet

Implementaciju robusnih ocenjivača na teroriji kredibiliteta ilustrujemo na jednodimenzionalnom slučaju multidimenzonalnog kredibiliteta.

Problem kredibiliteta posmatramo iz ugla čiste robusnosti, bez ikakvih odsecanja ekstremnih autlajera. Kao robusnu alternativa predlaže se ocenjivač

$$B_j^{M^a} = \beta^M + Z_j^M [B_j^M - \beta^M]$$

gde je $\beta^M = E[B_j^M]$, a B_j^M je *M*-ocenjivač nepoznatog parametra $\beta(\Theta_j)$. Budući da je matrica izloženosti riziku, W_j, unapred poznata bez umanjenja opštosti pretpostavlja se $W_j = I$ za svako j.

Ova pretpostavka proizilazi iz činjenice da se za poznatu matricu težina W_j problem težinskih najmanjih kvadrata može svesti na problem najmanjih kvadrata sledećom transformacijom $Y_j^* = W_j^{\frac{1}{2}}, X_j^* = W_j^{\frac{1}{2}} X_j$ i $\epsilon^* = W_j^{\frac{1}{2}} \epsilon$.

Dakle za svaki ugovor $j = 1, ..., K$ ocenjivač parametra β_j, odnosno B_j^M dobijamo kao rešenje problema

$$(\beta_j, S_j) = \arg\min \Big[\sum_{l=1}^{n} \left[\rho(\frac{Y_{ij} - x_{ij}^T \beta_j}{S_j}) + A \right] S_j \Big].$$

Teorema 37. *Pod pretpostavkama navedenim u prethodnom poglavlju ocenjivač kredibiliteta dat je sa*

$$Z_j^M = A^M(A^M + \Lambda_j)^{-1}$$

gde je $A^M = Cov[\beta_j^M(\Theta_j), \beta_j^M(\Theta_j)^T]$, $\Lambda_j = E[Cov[B_j^M, B_j^{M^T}]]$ *i* $\beta_j^M(\Theta_j) = E[B_j^M|\Theta_j]$.

Prvo je potrebno oceniti uticajnu funkciju

$$\hat{IF}[y_{ij}^T, X_{ij}, T^H, F_n^{(j)}] = n(X_j^T X_j)^{-1} \frac{\psi(r_{ij}^H/S_j)S_j}{n^{-1}\sum_{k=1}^n \psi'(r_{kj}^H)/S_j)} x_{ij} \tag{4.4}$$

gde je

$$r_{ij}^H = Y_{ij} - x_{ij}^T B_j^M,$$

a B_j^M je jedan *M* ocenjivač.

Empirijski ocenjivač kovarijanse $Cov[B_j^M, B_j^{M^T}|\Theta_j]$ može biti dobijen kao

$$\begin{aligned} \hat{v_j} &\backsimeq \frac{1}{n^2(n-p-1)} \sum_{i=1}^n \hat{IF}[x_{ij}^T, Y_{ij}, T^H, F_n^{(j)}] \times \hat{IF}[x_{ij}^T, Y_{ij}, T^H, F_n^{(j)}] \\ &= \frac{1}{(n-p-1)} \sum_{i=1}^n \Big[\frac{\psi^2(r_{ij}^H/S_j)S_j}{[\frac{1}{n}\sum_{k=1}^n \psi'(r_{kj}^H/S_j)]^2} \Big] (X_j^T X_j)^{-1} \\ &= \frac{1}{(n-p-1)} \sum_{i=1}^n (r_{\psi_{ij}^H})^2 (X_j^T X_j)^{-1} = \hat{s}_{M_j}^2 (X_j^T X_j)^{-1}, \end{aligned} \tag{4.5}$$

gde je

$$r_{\psi_{ij}}^H = \frac{\psi(r_{ij}^H/S_j)S_j}{\sum_{k=1}^n \psi'(r_{kj}^H/S_j)},$$

a $\hat{s}_{M_j}^2$ označava ocenjenu varijansu.

Empirijski ocenjivač parametra $\Lambda_j = E[Cov[B_j^M, B_j^M|\Theta_j]$ dat je sa

$$\hat{\Lambda}_j^* \simeq \hat{s}_M^{*^2} (X_j^T X_j)^{-1}, \tag{4.6}$$

gde je

$$\hat{s}_M^{*^2} = \frac{1}{K} \sum_{i=1}^K \hat{s}_{M_j}^2.$$

Da bi dobili nepristrasan ocenjivač $\Lambda_j = E[Cov[B_j^M, B_j^M|\Theta_j]$ svaki ugovor se mora korigovati faktorom L_j datim sa

$$L_j = 1 + \frac{p}{n} \Big[\frac{Var[\psi'(r_{ij}^H/S_j)]}{E[\psi'(r_{ij}^H/S_j)]^2} \Big],$$

gde $E[\psi'(r_{ij}^H/S_j)]$ ocenjujemo sa

$$E[\psi'(r_{ij}^H/S_j)] \approx \frac{1}{n} \sum_{i=1}^n \psi'(r_{ij}^H/S_j) =: m^{(j)}$$

i $Var[\psi'(r_{ij}^H/S_j)]$ ocenjujemo sa

$$Var[\psi'(r_{ij}^H/S_j)] \simeq \frac{1}{n}\sum_{i=1}^{n}[\psi'(r_{ij}^H/S_j) - m^{(j)}]^2.$$

Tada je empirijska ocenjivač Λ_j dat sa

$$\hat{\Lambda}_j \backsimeq \hat{s}_M^2(X_j^T X_j)^{-1},$$

gde je

$$s_M^2 = \frac{1}{K}\sum_{j=1}^{K} L_j \hat{s}_{M_j}^2.$$

Ocenjivač parametra A^M je dat sa

$$\hat{A}^M = \Big[\hat{u} - \frac{1}{K-1}\hat{s}_M^2(\sum_{j=1}^{K} X_j^T X_j)^{-1}\Big],$$

gde je

$$\hat{u} = \frac{1}{K-1}\sum_{j=1}^{K}[B_j^M - \hat{b}^M][B_j^M - \hat{b}^M]^T.$$

Empirijska ocena b^M data je sa

$$\hat{b}^M = \frac{1}{K}\sum_{j=1}^{K} B_j^M,$$

a ocena b data je sa $\hat{b} = \frac{1}{K}\sum_{j=1}^{K} B_j$.

Empirijski robusni regresioni kredibilitet dat je sa

$$\hat{B}_j^a \simeq \hat{b} + \hat{Z}_j^M[B_j^M - \hat{b}^M],$$

gde je

$$\hat{Z}_j^M = \hat{A}^M(\hat{A}^M + \hat{\Lambda}_j)^{-1}.$$

Zaključak

Ovaj rad prestavlja uvod u teoriju robusne statistike i implementaciju ovih metoda u teoriju kredibiliteta. Kroz rad smo stekli utisak o osetljivosti statistika na devijaciju od pretpostavki modela kojim ocenjujemo parametar. Kao alternativu klasičnim metodama za tačkastu ocenu parametara predlažemo robusnosne metode i robusne statistike. Za regresioni model, generalizovana metoda ocenjivača maksimalne verodostojnosti, odnosno klasa *M* ocenjivača, nije robusna na tačke jakog uticaja. Stoga uvodimo jednu robusnu klasu *S* ocenjivača i jednu finiju klasu *MM* ocenjivača koja poseduje visoku tačku preloma i visoko je efikasna.

Kada je reč o pravljenju modela, robusne regresione metode generalno ne trebaju biti korišćene za formiranje finalnog modela, sem u nekim okolnostima koje to dozvoljavaju. Ove metode mogu poslužiti i kao moćan alat za detekciju observacija koje pogoršavaju klasične ocenjivače metode *LS*. Prilikom analize podataka optimalna je upotreba više različitih robusnih ocenjivača i njihovo poređenje sa klasičnim (nerobusnim) ocenjivačima.

Osetljivost na ekstremna potraživanja (autlajere) i nedostatak robusnosti klasičnih metoda regresionih ocena kredibiliteta premije navela nas je na izučavanje robusne alternative. U Glavi 4. predstavljamo jedan model preciznih i robusnish ocenjivača krediblilteta konkretnog nosioca ugovora koji pripada datoj riziko klasi. Ovo je samo jedan od predloga za robustifikaciju ocenjivača kredibiliteta. Pored klase *MM* ocenjivača, predlažemo i druge, više robusne klase ocenjivača poput *LTS* i *LMS*. Pored ovog pristupa napominjemo da postoje i drugi pristupi koji vrše robustifikaciju premije.

Dodatak

Za generisanje Slika 3.2 i 3.5 korišćena je normalna raspodela $\mathcal{N}(0,1)$, a potom se biran prost slučajan uzorak sa x ose koji je pritom zamenjen Košijevom raspodelom $\mathcal{C}(0,40)$. U prilogu navodimo kod kojim je generisana Slika 3.5. Sličnim kodom generisane su i ostale slike.

```
x<-seq(1,20,length=20)
y<-c(5*x+7+rnorm(20,0,1))
s1<-sample(20,10)
for(j in s1){
  x[j]<-x[j]+rcauchy(1, location = 0, scale = 40)
}
plot(x[s1],y[s1],col="purple", pch=20, xlab="x", ylab="y")
points(x[-s1],y[-s1],pch=20)

rlmH <- rlm(y~x, method = "MM", psi = psi.bisquare,
                                    c = 4.685, maxit = 60)
abline(rlmH, col="red", lwd=2)

rlmL <- rlm(y~x, method = "MM", psi = psi.bisquare,
                                    c = 2.973, maxit = 60)
abline(rlmL, col="blue", lwd=2)

legend(-320, 80, '75.9% efikasan MM', box.lwd = 2, box.col = "blue")
legend(-320, 95, '95.7% efikasan MM', box.lwd = 2, box.col = "red")
```

Bibliografija

[1] Alamgir, Amjad Ali, Sajjad Ahmad Khan, Dost Muhammad Khan and Umair Khalil *A New Efficient Redescending M- Estimator: Alamgir Redescending M- estimator*. Research Journal of Recent Sciences, Vol. 2(8), 79-91, August, (2013).

[2] Ana M. Bianco, Marta Garcia Ben and Víctor J. Yohai *Robust Estimation for Linear Regression with Asymmetric Errors*. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, Vol. 33, No. 4(Dec., 2005), pp. 511-528.

[3] Brenton R. Clarke *Nonsmooth Analysis and Fréchet Differentiability of M-Functionals*. Probability Theory and Related Fields 73(2):197-209, September (1986).

[4] Brenton R. Clarke *Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations*. Annals of Statistics, Volume 11, Number 4 (1983), 1196-1205.

[5] C. G. Small, J. Wang & Z. Yang *Eliminating multiple root problems in estimation*. Statistical Science, 15, 313-341,(2000).

[6] Chi-Lun Cheng *Robust Linear Regression via Bounded Influence M-estimators* Journal Of Mmultivariate Analysis 40, 158-171, (1992).

[7] David A. Belsley, Edwin Kuh, Roy E. Welsch *Regression Diagnostics Identifying Influential Data and Sources of Collinearity*. Wiley-Interscience, (2004).

[8] David G. Luenberger, Yinyu Ye *Linear and Nonlinear Programming*. Springer, (2015).

[9] De Vylder, F. *Parameter estimation in credibility theory* ASTIN Bulletin, 10 (1978), pp 99-112.
doi:10.1017/S0515036100006395

[10] Douglas P. Wiens, Eden K.H. Wu *A comparative study of robust designs for M-estimated regression models*. Computational Statistics and Data Analysis 54, 1683-1695. (2010).

[11] E.L. Lehmann *Theory of Point Estimation*. Springer, (1983).

[12] Ezequiel Smucler *Asymptotic Statistical Properties of Redescending M-estimators in Linear Models with Increasing Dimension*. Instituto de Calculo, Universidad de Buenos Aires, (2016).

arXiv:1612.05951

[13] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, Werner A. Stahelauth *Robust Statistics The Approach Based on Influence Functions*. John Wiley & Sons, 1986.

[14] Georgios Pitselis *A review on robust estimators applied to regression credibility*. Journal of Computational and Applied Mathematics Volume 239, 1 February 2013, Pages 231–249.

[15] Georgios Pitselis *Application of GM and MM Estimators to Regression Credibility*.

[16] Georgios Pitselis *Robust regression credibility: The influence function approach* Insurance: Mathematics and Economics Volume 42, Issue 1, February 2008, Pages 288–300.

[17] Georgy Shevlyakov, Stephan Morgenthaler, Alexander Shurygin *Redescending M-estimators*. Journal of Statistical Planning and Inference 138(10), (2008).

DOI: 10.1016/j.jspi.2007.11.008

[18] Hans Bühlmann, Alois Gisler *A Course in Credibility Theory and its Applications* Springer (2005).

[19] Helmut Rieder *Robust Asymptotic Statistics*. Springer, (1994).

[20] Jana Jurečková, Pranab Kumar Sen, Jan Picek *Methodology in robust and nonparametric statistics*. CRC Press, 2013.

[21] M. Kurilić *Osnovi opšte topologije*. Univerzitet u Novom Sadu, Prirodno-matematički fakultet, (1998).

[22] Michael B. Dollinger & Robert G. Staude *Influence Functions of Iteratively Reweighted Least Squares Estimators* Journal of the American Statistical Association, 86:415, 709-716, (1991).

[23] Michael B. Dollinger And Robert G. Staude *The Construction Of Equileverage Designs For Multiple Linear Regression*. Austral. J. Statist., 32(1), (1990), 99-118.

DOI: 10.1111/j.1467-842X.1990.tb01002.x

[24] Nassim Nicholas Taleb *Antifragile: Things That Gain from Disorder*. Random House, (2012).

[25] Norman R. Draper, Harry Smith *Applied Regression Analysis, Third Edition*. Wiley-Interscience, (1998).

[26] Peter J. Huber, Elvezio M. Ronchetti *Robust statistics*. Wiley, (2009).

[27] Peter J. Rousseeuw, Annick M. Leroy *Robust Regression and Outlier Detection* Wiley, (1987).

[28] Philippe G. Ciarlet *Linear and Nonlinear Functional Analysis with Applications*. SIAM Society for Industrial and Applied Mathematics, (2013).

[29] Ricardo A. Maronna, Douglas R. Martin, Victor J. Yohai *Robust Statistics, Theory and Methods*. Wiley, (2006).

[30] Ricardo A. Maronna, Victor J. Yohai *Asymptotic Behavior of General M-estimates for Regression and Scale with Random Carriers*. Springer-Verlag, 1981.

[31] Ricardo A. Maronna, Victor J. Yohai *Correcting MM estimates for "fat" data sets*. Computational Statistics and Data Analysis 54 3168-3173, (2010).

[32] Robert G. Staudte, Simon J. Sheather *Estimation and Testing*. John Wiley & Sons, 1990.

[33] Robert J. Serfling *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc, 1980.

[34] Sanford Weisberg *Applied Linear Regression*. Wiley, (2013).

[35] Stanković, B., Pilipović, S. *Teorija distribucija*. Publ. Math. Inst. Novi Sad, Novi Sad, (1983).

[36] Stevan Pilipovic, Dora Selesi *Mera i integral*. ZAVOD ZA UDŽBENIKE, Beograd, 2012.

[37] Stuart A. Klugman, Harry H. Panjer & Gordone E. Wilmot *Loss Models From Data to Decision 2nd ed*. Wiley, (2004).

[38] Victor J. Yohai *High Breakdown point and high efficiency estimator for robust regression*. The annals of Statistics, Vol. 15, No. 20, 642-656. (1987).

[39] Victor Y. Yohai, Ricardno A. Maronna *Asymptotic Behavior of M-Estimators for the Linear Model*. The Annals of Statistics, Vol. 7, No. 2, 258-268, (1979).

[40] William H. Greene *Econometric Analysis*. Prentice Hall, (2011).

[41] X. R. Chen And Y. H. Wu *Strong Consistency of M-Estimates in Linear Models*. Journal Of Mmultivariate Analysis 27, 116-130, (1988).

[42] Yuliana Susanti, Hasih Pratiwi, Sri Sulistijowati H, Twenty Liana *M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION*. International Journal of Pure and Applied Mathematics Volume 91 No. 3, 349-360 (2014).

Kratka biografija

Mihailo Tomišić je rođen 05. septembra 1991. godine u Somboru. Osnovnu školu "Vuk Stefanović Karadžić" u Crvenki završio je 2006. godine. Potom upisuje gimnazuju "Žarko Zrenjanin" u Vrbasu prirodno matematički smer koju završava 2010. godine. Iste godine upisuje Prirodno-matematički fakultet, smer Primenjena matematika, modul Matematika finansija, gde 2014. godine završava osnovne studije. Iste godine upisuje i master studije na Prirodno-matematičkom fakultetu. Zaključno sa semtembarskim ispitnim rokom polaže sve ispite predviđene nastavnim planom i programom master studija, čime stiče uslov za odbranu ovog master rada.

UNIVERZITET U NOVOM SADU
PRIRODNO-MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR
Identifikacioni broj:
IBR
Tip dokumentacije: Monografska dokumentacija
TD
Tip zapisa: Tekstualni štampani materijal
TZ
Vrsta rada: Master rad
VR
Autor: Mihailo Tomišić
AU
Mentor: Prof. dr Dora Seleši
MN
Naslov rada: Robusne statistike i ocenjivači maksimalne verodostojnosti i njihova primena u teoriji kredibiliteta
NR
Jezik publikacije: srpski (latinica)
JP
Jezik izvoda: srpski/engleski
JI
Zemlja publikovanja: Republika Srbija
ZP
Uže geografsko područje: Vojvodina
UGP
Godina: 2017.
GO
Izdavač: Autorski reprint
IZ
Mesto i adresa: Novi Sad, Prirodno-matematički fakultet, Trg Dositeja Obradovića 4
MA
Fizički opis rada: 4/79/42/7/1/19/1
(broj poglavlja/strana/lit. citata/tabela/slika/grafika/priloga)
FO
Naučna oblast: matematika
NO
Naučna disciplina: primenjena matematika
ND
Predmetna odrednica/Ključne reči: tačkasti robusni ocenjivači, kvalitativna i kvantitativna robusnost, tačka preloma, uticajna funkcija, B-robusne statistike, M ocenjivači, Hachmeisterov model kredibiliteta, robusni model kredibiliteta.
PO

UDK:
Čuva se: u biblioteci Departmana za matematiku i informatiku, Novi Sad
ČU
Važna napomena:
VN
Izvod: Cilj rada je analiza robusnosti klasičnih (nerobusnih) statistika i prezentovanje robusne klase ocenjivača maksimalne verodostojnosti, odnosno M ocenjivača. Prezentovani su statistike za višeparametarsku ocenu lokacije, skale i regresije. Pokazana je neograničenost uticajne funkcije ovih statistika. Osetljivost klasičnih metoda na odstupanje od pretpostavki modela, odnosno outlajere i tačke jakog uticaja iskazuje jasnu potrebu za robusnijim metoda za ocenu parametara kredibilitta.
IZ

Datum prihvatanja teme od strane NN Veća: 08.05.2017.
DP
Datum odbrane:
DO
Članovi komisije:
Predsednik: dr Danijela Rajter-Ćirić, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Dora Seleši, vanredni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Nataša Krklec- Jerenkić, docent, Prirodno-matematički fakultet, Univerzitet u Novom Sadu
KO

UNIVERSITY OF NOVI SAD
FACULTY OF NATURAL SCIENCES AND MATHEMATICS
KEY WORDS DOCUMENTATION

Accession number:
ANO
Identification number:
INO
Document type: Monograph type
DT
Type of record: Printed text
TR
Contents code: Master thesis
CC
Author: Tomišić Mihailo
AU
Mentor: Dora Seleši, PhD
MN
Title: Robust statistics and maximum likelihood estimators and its application in credibility theory
TI
Language of text: Serbian
LT
Language of abstract: Serbian/English
LA
Country of publication: Serbia and Montenegro
CP
Locality of publication: Vojvodina
LP
Publication year: 2017.
PY
Publisher: Author's reprint
PU
Publication place: Novi Sad, Faculty of Science and Mathematics, Dositeja Obradovića 4
PP
Physical description: 4/79/42/7/1/19/1
(chapters/pages/literature/tables/pictures/graphics/appendices)
PD
Scientific field: Mathematics
SF
Scientific discipline: applied mathematics
SD
Subject / Key words: robust point estimators, qualitative and quantitative robustness, breaking point, influence function, B-robust statistics, M estimators, Hachmeisterov's model, robust credibility model.
SKW

UC:
Holding data: library of the Department of Mathematics and Informatics, Novi Sad
HD
Note:
N
Abstract: The goal of this paper is analysis of robustness of classical (non robust) statistics and presentation of robust class maximum likelihood type of estimators, M estimators. Presented methods include multivariate estimation of location and regression models. It was demonstrated unboundedness of influence function of these statistics. The sensitivity of classical methods on departure from the model assumptions, outliers and leverage points shows a clear need for a more robust method of estimating the parameters credibility premium.
AB

Thesis defend board:
President: Prof. Danijela Rajter-Ćirić, PhD, Faculty of Science and Mathematics, University in Novi Sad

Member: Prof. Dora Seleši, PhD,Faculty of Science and Mathematics, University in Novi Sad

Member: Assist. prof. Nataša Krklec- Jerenkić, PhD, Faculty of Science and Mathematics, University in Novi Sad
DB