LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

EMPĪRISKĀ TICAMĪBAS FUNKCIJA AR NOVĒRTĒTIEM PARAMETRIEM

KURSA DARBS

Autors: **Leonora Pahirko**

Stud. apl. lp06061

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2011

Saturs

Ievads

Statistikā daudzas procedūras balstās uz parametriskiem pieņēmumiem par datu sadalījumu, taču praksē bieži nākas sastapties ar situāciju, kad ir grūti noteikt sadalījuma likumu, kam par iemeslu var būt nepietiekošs izlases apjoms vai arī sarežģīts teorētiskais sadalījums. Tāpēc aizvien populārākas kļūst neparametriskās metodes, lai izvairītos no kļūdām, kas rodas, lietojot neprecīzus pieņēmumus par datu sadalījumu. Viena no tādām metodēm ir empīriskās ticamības (turpmāk EL) funkcijas metode, kas ļauj modelēt nezināmo sadalījumu, balstoties uz datu kopu.

Owen ([1],[2]) ir viens no pirmajiem, kas sāka pielietot EL metodi ticamības intervālu un reģionu konstruēšanai, taču EL metodes pirmsākumi meklējami Thomas, Grunkemeier (1975, [3]) darbos par ticamības intervālu novērtēšanu izdzīvošanas datu analīzē. Tomēr Owen vispārināja rezultātus, kas bija iegūti parametriskai vislielākās ticamības metodei, un parādīja, ka EL metodes statistika tiecas uz χ^2_p sadalījumu.

Šī darba mērķis ir iepazīties ar rezultātiem, kas vispārina EL metodes pielietošanu gadījumos, kad tiek novērtēti traucējošie (*nuisance*) parametri ar kādu no neparametriskajām metodēm. Šī problemātika smalki iztirzāta Hjort, McKeague un Van Keilegom publikācijā [4]. Autori arī apskata vairākus piemērus, kuros EL metodes pielietošana ir ērtāka nekā citas metodes.

Arī traucējošo parametru novērtēšana (*plug-in*) EL metodes ticamības reģionu konstruēšanā nav jauna. Pēdējā laikā tā tikusi bieži pielietota dažādos izdzīvošanas datu analīzes kontekstos, piemēram Wang un Jing (2001, [5]), kā arī izlases apsekojumos ar trūkstošo datu imputāciju. Tomēr autoru mērķis darbā [4] ir paplašināt pielietošanas iespējas EL metodei ar *plug-in* parametriem, kas varētu nodrošināt arī darbības ar plašām piemēru matricām. Lai to izdarītu, tiek ieviesti vispārēji pieņēmumi nosacījumu veidā, kurus ir viegli pārbaudīt lielākajai daļai EL metodes pielietojumos, kas ir saprātīgāk nekā ieviest teorētiskos pamatelementus katram atsevišķam gadījumam.

Šī darba uzdevumi ir iepazīties ar pamatdefinīcijām un teorēmām par El metodi, iepazīties ar EL metodi ar novērtētiem parametriem un ar simulāciju palīdzību pārbaudīt, kā praktiski strādā piedāvātais teorētiskais materiāls. Darbs sastāv no 4 nodaļām un 1 pielikuma. 1. nodaļa veltīta empīriskai ticamības funkcijai un empīriskās ticamības funkcijas attiecības testam, 2. nodaļā apskatīta ticamības intervālu konstruēšana ar EL metodi vidējai vērtībai un sniegts īss ieskats ticamības intervālu konstruēšanai vispārējā

gadījumā. 3. nodaļa apraksta EL metodi ar novērtētiem parametriem, galvenos rezultātus un to pierādījumu, kas pārstrādāts uz viendimensionālu gadījumu. 4. nodaļā doti daži EL metodes ar novērtētiem parametriem pielietojumi, kas apskatīti gan no teorētiskā, gan praktiskā aspekta. Pielikumā iekļauts programmas R kods simulāciju veikšanai.

1. Neparametriskās ticamības funkcijas attiecības tests

1.1. Neparametriskā vislielākās ticamības funkcija

Definīcija 1. $X_1,\ldots,X_n$ *iid*, empīriskā sadalījuma funkcija tiek definēta kā

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

kur $-\infty < x < \infty$ un indikatorfunkcija

$$I_{\{X_i \leq x\}} = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}.$$

Definīcija 2. $X_1,\ldots,X_n \sim F$ *iid*, funkcijas F neparametriskā (empīriskā) ticamības funkcija ir

$$L(F) = \prod_{i=1}^{n} p_i,$$

kur $p_i = P(x = X_i)$.

Ja F ir nepārtraukts sadalījums, tad $L(F) = 0$. Lai empīriskā ticamības funkcija būtu pozitīva, sadalījuma funkcijai F jāuzliek pozitīva varbūtība katram no pētāmās datu kopas punktiem.

Teorēma 1. *(Owen,[2]) Pieņemsim, ka* $X_1,\ldots,X_n \in \mathbb{R}$ *ir neatkarīgi gadījuma lielumi ar sadalījuma funkciju* F*, un* $\widehat{F}_n$ *ir to empīriskā sadalījuma funkcija. Ja* $F \neq \widehat{F}_n$*, tad* $L(F) < L(\widehat{F}_n)$.

Teorēma 1. pierāda, ka neparametrisko ticamības funkciju maksimizē empīriskā sadalījuma funkcija. Tātad empīriskā sadalījuma funkcija $\widehat{F}_n$ ir funkcijas F neparametriskās vislielākās ticamības funkcijas novērtējums.

1.2. Empīriskās ticamības funkcijas attiecības tests

Neparametriskās ticamības funkcijas attiecību var izmatot hipotēžu pārbaudei un ticamības intervālu konstruēšanai, līdzīgi kā parametrisko vislielākās ticamības attiecības testu (sk. [6]).

Definēsim neparametriskās ticamības funkcijas attiecību sadalījuma funkcijai F ar izteiksmi

$$R(F) = \frac{L(F)}{L(\widehat{F}_n)} = \prod_{i=1}^{n} np_i.$$

Lai novērtētu parametrus un veiktu hipotēžu pārbaudi, nezināmo parametru θ izsaka kā funkcionāli no sadalījuma funkcijas F, t.i., $\theta = T(F)$ un θ novērtējums ir $\hat{\theta} = T(\widehat{F}_n)$.

Uzskatīsim, ka F pieder kopai $\mathcal{F}$. $\mathcal{F}$ var sakrist ar $\mathbb{R}$, taču biežāk tiek izmantota kāda $\mathbb{R}$ apakškopa. Definēsim profila empīriskās ticamības attiecības funkciju

$$\mathrm{EL}(\theta) = \max\{R(F)|T(F) = \theta, \quad F \in \mathcal{F}\}.$$

Empīriskās ticamības hipotēžu pārbaude noraida $H_0 : T(F_0) = \theta_0$, kad $\mathrm{EL}(\theta_0) < r_0$, kur r_0 ir konstante, kuru var noteikt izmantojot 2.1. nodaļā apskatītos rezultātus (sk. arī [7]). Empīriskās ticamības metodes ticamības reģioni ir formā $\{\theta|\mathrm{EL}(\theta) \geq r_0\}$.

2. Ticamības intervālu konstruēšana ar EL

Lai labāk izprastu ticamības intervālu konstruēšanas problemātiku, apskatīsim vienkāršāko gadījumu $\theta = T(F) = \mu$.

Pieņemsim, ka mums ir doti $X_1,\ldots,X_n$ *iid* gadījuma lielumi ar nezināmu sadalījuma funkciju F. Lai konstruētu ticamības reģionus parametram $\mu = EX = \int_{-\infty}^{+\infty} xdF(x)$, izmantosim profila empīriskās ticamības attiecības funkciju

$$\mathrm{EL}(\mu) = \max\left\{\prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu\right\}. \tag{2.1}$$

Owen ([2]) parādīja, ka eksistē viens vienīgs izteiksmes (2.1) labās puses atrisinājums, ja μ atrodas izliektajā čaulā, ko veido $X_1,\ldots,X_n$. Maksimizācijas problēmu (2.1) var atrisināt ar Lagranža reizinātāju palīdzību.

Izteiksme $\prod_{i=1}^{n} np_i$ pie ierobežojumiem

$$p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu$$

savu maksimumu sasniedz, kad

$$p_i = p_i(\mu) = n^{-1}\{1 + \lambda(X_i - \mu)\}^{-1},$$

kur $\lambda = \lambda(\mu)$ var iegūt no izteiksmes

$$\sum_{i=1}^{n} \{1 + \lambda(X_i - \mu)\}^{-1}(X_i - \mu) = 0.$$

Tātad

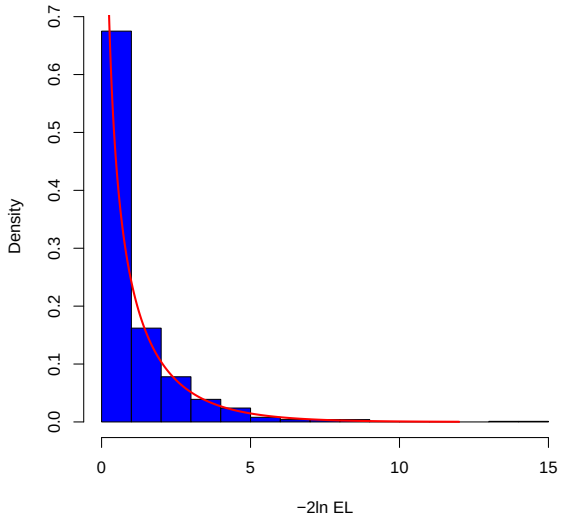
$$\mathrm{EL}(\mu) = \prod_{i=1}^{n} \{1 + \lambda(X_i - \mu)\}^{-1}.$$

Logaritmiskā empīriskās ticamības attiecības statistika ir $-2\ln \mathrm{EL}(\mu)$, t.i,

$$W_E(\mu_0) = -2\ln \mathrm{EL}(\mu) = 2\sum_{i=1}^{n} \ln\{1 + \lambda(X_i - \mu)\}.$$

Piezīme 2. *Lagranža reizinātājs* λ *meklējams intervālā* $\dfrac{1 - n^{-1}}{\mu - X_{(n)}} < \lambda(\mu) < \dfrac{1 - n^{-1}}{\mu - X_{(1)}}$, *kur* $X_{(1)},\ldots,X_{(n)}$ *ir augošā secībā sakārtota sākotnējā izlase [2].*

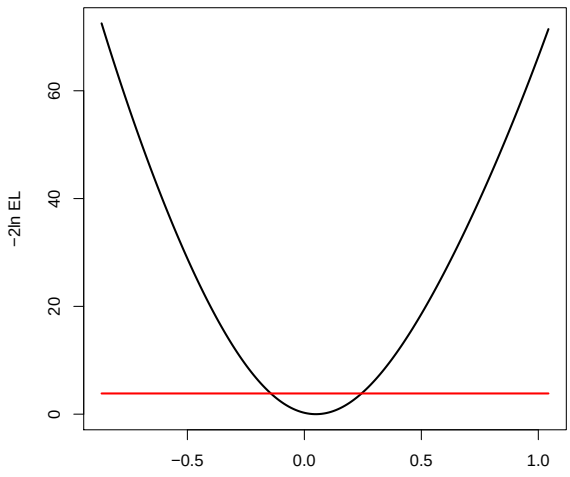
Teorēma 3. *(Owen, [2]) Ja* $X_1,\ldots,X_n$ *iid gadījuma lielumi ar sadalījuma funkciju* F *un* $\mu_0 = E(X_i)$, *un* $0 < D(X_i) < \infty$, *tad* $-2\ln EL(\mu_0)$ *tiecas uz* χ_1^2, *kad* $n \to \infty$.



1. att. Histogramma $-2\ln \mathrm{EL}(\mu_0)$ un χ^2_1 blīvuma funkcija

Ar simulāciju palīdzību tika pārbaudīta Teorēma 3, un iegūtie rezultāti aplūkojami 1. attēlā. Redzams, ka χ^2_1 blīvuma funkcija labi aproksimē $-2\ln \mathrm{EL}(0)$ 1000 reizes ģenerētiem datiem ar sadalījumu $N(0,1)$.

Kā izklāstīts Teorēmā 3, ja $\mu = \mu_0$, tad $W_E(\mu_0)$ tiecas uz χ^2_1, kad $n \to \infty$. Ticamības intervālus parametram μ pie nozīmības līmeņa α var iegūt kā tādu punktu kopu, kuriem $W_E(\mu_0) \leq c$, kur c var izteikt no $P(\chi^2_1 \leq c) = 1 - \alpha$.



2. att. Ticamības intervāli vidējai vērtībai ar EL metodi $N(0,1)$

1. tabula Pārklājumu precizitāte ticamības intervāliem vidējai vērtībai

	$n = 20$	$n = 50$	$n = 100$
EL	0.9318	0.9524	0.949
t.test	0.9501	0.9513	0.9469

2. attēlā redzams ticamības intervāls vidējai vērtībai $\mu_0 = 0$ standartnormāli sadalītiem simulētiem datiem, $\alpha = 0.05$. Tā kā īstā vērtība atrodas ticamības intervāla robežās, kas ir -0.1432 un 0.2445, tad varam domāt, ka EL metode strādā labi, bet, lai par to pārliecinātos, aplūkosim pārklājumu precizitāti ticamības intervāliem. Tabulā 1. attēlota gan EL metodes, gan $t-$testa ticamības intervālu pārklājumu precizitāte izlases apjomiem $n = 20$, 50 un 100. Redzams, ka pie nelieliem izlašu apjomiem EL metode strādā nedaudz sliktāk, par t-testu, taču 10 000 reižu simulētiem datiem ar $n = 50$ un 100 pārklājumu precizitāte abiem testiem ir līdzīga, tātad kopumā varam secināt, ka EL metode strādā labi.

2.1. EL un ticamības intervāli vispārējā gadījumā

Empīriskās ticamības funkcijas metodes vispārējs gadījums izklāstīts materiālā [7]. Mēs apskatīsim tikai galvenos rezultātus.

Pieņemsim, ka mums ir doti neatkarīgi un vienādi sadalīti p-dimensionāli gadījuma lielumi X_1,...,X_n ar nezināmu sadalījuma funkciju F. Informācija par F un d-dimensionālu parametru θ ir dota funkcionāli – p neatkarīgu, nenovirzītu funkciju formā, t.i., $m_j(X, \theta)$, $j = 1, 2, \ldots, p$, turklāt tādas, ka $E\{m_j(X, \theta)\} = 0$.

Piezīme 4. *Turpmāk šīs funkcijas $m_j(X, \theta)$ sauksim par novērtējošām funkcijām, un p-dimensionālā gadījumā apzīmēsim tās ar novērtējošo funkciju vektoru $m(X, \theta)$.*

Tātad neparametriskai ticamības funkcijai $L(F) = \prod_{i=1}^{n} p_i$ jāatrod maksimums pie ierobežojumiem

$$p_i > 0, \quad \sum_i p_i = 1, \quad \sum_i p_i m(X_i, \theta) = 0.$$

Lai to izdarītu, pielietosim Lagranža reizinātāju metodi tāpat kā vidējās vērtības gadījumā, un varam iegūt

$$p_i = \frac{1}{n(1 + \lambda^T m(X_i, \theta))}.$$

Tagad varam definēt arī logaritmisko profila empīriskās ticamības attiecības funkciju

$$\mathrm{el}(\theta) = \sum_{i=1}^{n} \ln\{1 + \lambda^T(\theta) m(X_i, \theta)\}. \tag{2.2}$$

Piezīme 5. *Gadījumā, ja* $\theta = \mu$, *tad novērtējošā funkcija* $m(X_i, \mu) = X_i - \mu$ *un* $2^{-1}W_E(\mu)$ *ir (2.2) speciālgadījums.*

Qin un Lawles ([7]) savā darbā apskata nosacījumus, kuriem izpildoties, var pierādīt, ka empīriskās ticamības attiecības statistika hipotēzei $H_0 : \theta = \theta_0$ ir

$$W_E(\theta_0) = 2\mathrm{el}(\theta_0) - 2\mathrm{el}(\widehat{\theta}),$$

kur $\widehat{\theta}$ ir parametra θ empīriskās vislielākās ticamības novērtējums. Turklāt, $W_E(\theta_0) \to \chi^2_p$, kad $n \to \infty$ un H_0 ir spēkā.

Šie rezultāti dod iespēju konstruēt ticamības intervālus parametram θ tāpat kā vidējai vērtībai.

3. Empīriskās ticamības funkcijas metode ar novērtētiem parametriem

Kā jau apskatījām iepriekšējās nodaļās, EL metode ļauj modelēt nezināmo sadalījumu, izmantojot dotos novērojumus. Secinājumi par mūs interesējošajiem parametriem var tikt izdarīti, lietojot p-dimensionālu novērtējošo funkciju formā $m_n(X, \theta, h)$, kur h ir "traucējošais" (*nuisance*) parametrs ar nezināmu īsto vērtību h_0.

Kad h_0 ir zināms, mēs varam aizstāt h ar h_0 profila EL attiecības funkcijā

$$\mathrm{EL}_n(\theta, h) = \max \left\{ \prod_{i=1}^{n} np_i \mid p_i > 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m_n(X_i, \theta, h) = 0 \right\},$$

un atrast ticamības intervālus parametram θ_0 formā $\{\theta : \mathrm{EL}_n(\theta, h_0) > c\}$, kur problemātika, kā atrast c apskatīta [1].

Autori N.L. Hjort, I.W. Mckeague un I. Van Keilegom publikācijā [4] apskata nosacījumus $(A0)$ - $(A3)$, kuriem izpildoties ir iespējams vispārināt Ovena EL teorēmu [1]. Nezināmais h_0 tiek aizstāts ar novērtējumu $\hat{h}$, un tiek pieļauts, ka novērtējošā funkcija var būt atkarīga no n. Turpmāk apskatītie nosacījumi $(A0)$ - $(A3)$ nodrošina nedeģenerētu robežsadalījumu un var būt noderīgi specifiskos pielietojumos, piemēram, kad novērojumi nav vienādi neatkarīgi sadalīti.

Ieviesīsim sekojošus apzīmējumus vektoriem v, $|v|$ - Eiklīda norma un $v^{\otimes 2} = vv^T$, un matricām $V = (v_{ij})$, $|V| = \max_{ij} |v_{ij}|$.

$\{a_n\}$ - pozitīvu konstanšu virkne un U - nedeģenerēts p-dimensionāls gadījuma vektors. V_2 - $p \times p$ pozitīvi definīta kovariāciju matrica.

Piezīme 6. *Ja nav norādīts citādāk, tad izmantosim* $a_n = 1$ *un* $U \sim N_p(0, V_1)$, *kur* V_1 *pozitīvi definīta kovariāciju matrica.*

(A0) $P(\mathrm{EL}_n(\theta_0, \hat{h}) = 0) \to 0$.

(A1) $\sum_{i=1}^{n} m_n(X_i, \theta_0, \hat{h}) \to_d U$.

(A2) $a_n \sum_{i=1}^{n} m_n^{\otimes 2}(X_i, \theta_0, \hat{h}) \to_{pr} V_2$.

(A3) $a_n \max_{1 \le i \le n} ||m_n(X_i, \theta_0, \hat{h})|| \to_{pr} 0$.

Teorēma 7. *Ja* $(A0)$ - $(A3)$ *ir spēkā, tad* $-2a_n^{-1} \log EL_n(\theta_0, \widehat{h}) \to_d U^t V_2^{-1} U$.

Piezīme 8. *Ovena EL teorēma seko no Teorēmas 7, izvēloties* $a_n = 1$ *un* $m_n(X_i, \theta_0, h) = m(X_i, \theta_0, h)/\sqrt{n}$.

3.1. Viendimensionāls gadījums

Vispirms apskatīsim, kā nosacījumi $(A0)$ - $(A3)$ izskatās viendimensionālā gadījumā, un pieņemsim, ka $m_n(X_i, \theta_0, h) = m(X_i, \theta_0, h)/\sqrt{n}$, kā arī U ir normāli sadalīts gadījuma lielums ar vidējo vērtību 0 un dispersiju σ^2. Tātad $V_2 = \sigma^2$.

Tad

(A0) $P(\mathrm{EL}_n(\theta_0, \hat{h}) = 0) \to 0$.

(A1) $\sum_{i=1}^n m(X_i, \theta_0, \hat{h})/\sqrt{n} \to_d U$.

(A2) $\sum_{i=1}^n m^2(X_i, \theta_0, \hat{h})/n \to_{pr} \sigma^2$.

(A3) $\max_{1\le i\le n} |m(X_i, \theta_0, \hat{h})\sqrt{n}| \to_{pr} 0$.

Un Teorēma 7 dos rezultātu formā $-2\log EL_n(\theta_0, \widehat{h}) \to_d U^2/\sigma^2$, kas sakrīt ar Owena EL teorēmas rezultātu.

Lai labāk izprastu nosacījumu $(A0)$ - $(A3)$ nepieciešamību, pārbaudīsim, vai tie izpildās vidējai vērtība, t.i., kad $m(X_i, \theta) = X_i - \mu$.

Tā kā pieņēmām, ka varbūtības p_i lielākas par 0, tad $(A0)$ vienmēr izpildās. $(A1)$ seko no Centrālās robežteorēmas [6]:

$$\sum_{i=1}^n m(X_i, \theta_0, \hat{h})/\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) = \frac{1}{\sqrt{n}} \left(\sum_{i=1}^n X_i - n\mu \right) = \sqrt{n}(\bar{X} - \mu) \to N(0, \sigma^2).$$

Kā redzams, arī $(A2)$ izpildās, izmantojot Lielo skaitļu likumu [6]:

$$\sum_{i=1}^n m^2(X_i, \theta_0, \hat{h})/n = \sum_{i=1}^n \frac{(X_i - \mu)^2}{n} = \widehat{\sigma}^2 \to_{pr} \sigma^2.$$

Savukārt nosacījums $(A3)$ ir spēkā, jo

$$\max_{1\le i\le n} |m(X_i, \theta_0, \hat{h})/\sqrt{n}| = \max_{1\le i\le n} |(X_i - \mu)/\sqrt{n}|, \tag{3.1}$$

un tā kā $\max_{1\le i\le n} |X_i| = o_{pr}(\sqrt{n})$ [1], tad (3.1) pēc varbūtības tiecas uz 0.

Redzams, ka dotie nosacījumi apraksta statistikas likumus, kas ir vienmēr spēkā *iid* novērojumiem ar galīgu dispersiju. Taču, kad tiek pielietoti *plug-in* novērtējumi, situācija var būt pretēja, un ir svarīgi, lai izpildās $(A0)$ - $(A3)$.

Pierādīsim, ka, izpildoties $(A0)$ - $(A3)$, $-2\log EL_n(\theta_0, \widehat{h}) \to_d U^2/\sigma^2$ viendimensionālā gadījumā.

Pierādījums. Apzīmēsim $X_{n,i} = m_n(X_i, \theta_0, \widehat{h})$.

No $(A0)$ $\mathrm{EL}_n = \mathrm{EL}_n(\theta_0, \widehat{h}) = \prod_{i=1}^n (1+\widehat{\lambda}X_{n,i})^{-1}$, kur Lagranža reizinātājs $\widehat{\lambda}$ apmierina vienādojumu $\sum_{i=1}^n X_{n,i}/(1+\widehat{\lambda}X_{n,i}) = 0$.

Tad EL statistiku varam izteikt formā

$$-2\ln \mathrm{EL}_n = G_n(\widehat{\lambda}) = \sup_{\lambda} G_n(\lambda), \tag{3.2}$$

kur $G_n(\lambda) = 2\sum_{i=1}^n \ln(1+\lambda X_{n,i})$, un G_n definīcijas apgabals ir kopa, kurā tā ir definēta (attiecībā uz $\ln x$, kas nav definēts $x \le 0$). Jāatzīmē, ka G_n ir ieliekta un sasniedz maksimumu pie $\widehat{\lambda}$, tā kā $dG_n(\widehat{\lambda})/d\widehat{\lambda} = 0$.

Tālāk apskatīsim G_n kvadrātisko aproksimāciju

$$G_n^*(\lambda) = 2\lambda U_n - \lambda^2 V_n, \quad \text{kur} \quad U_n = \sum_{i=1}^n X_{n,i}, \quad V_n = \sum_{i=1}^n X_{n,i}^2,$$

un G_n^* definīcijas apgabals ir $\mathbb{R}$.

Nedaudz vēlāk pierādīsim, ka starpība starp maksimālajām G_n un G_n^* vērtībām ir ar kārtu $o_{pr}(a_n)$.

Tad no (3.2) un no fakta, ka G_n^* tiek maksimizēta pie $\lambda^* = V_n^{-1}U_n$, kad V_n nav 0, seko, ka

$$-2a_n^{-1}\ln \mathrm{EL}_n = a_n^{-1}\sup_{\lambda} G_n^*(\lambda) + o_{pr}(1) = U_n^2(a_n V_n)^{-1} + o_{pr}(1), \tag{3.3}$$

kas pēc sadalījuma tiecas uz U^2/σ^2, ja pieņemam, ka $(A1)$ un $(A2)$ ir spēkā. No šī pierādījuma redzams, ka Teorēma 7 ir spēkā gadījumos, kad $(U_n, V_n) \to_d (U, \sigma^2)$.

Tagad pierādīsim, ka $\sup G_n - \sup G_n^* = o_{pr}(a_n)$.

Pirmkārt, noteiksim $\widehat{\lambda}$ stohastisko kārtu. Rakstīsim $\widehat{\lambda} = |\widehat{\lambda}|$, un, tāpat kā [2], ir spēkā nevienādība

$$|\widehat{\lambda}|(V_n - D_n U_n) \le U_n,$$

kur $D_n = \max_{i\le n}|X_{n,i}|$. Bet $V_n = o_{pr}(a_n^{-1})$, $U_n = o_{pr}(1)$ un $D_n U_n = o_{pr}(a_n^{-1})$, tāpēc $|\widehat{\lambda}| = o_{pr}(a_n)$. Turklāt $\lambda^* = V_n^{-1}U_n$, kad V_n nav 0, tāpēc λ^* ir ar tādu pašu stohastisko kārtu $o_{pr}(a_n)$ kā $\widehat{\lambda}$.

Ir zināms, ka $\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 h(x)$, kur $|h(x)| \le 2$ katram $|x| \le \frac{1}{2}$. Tas dod, katram $c > 0$ un $|\lambda| \le c$,

$$G_n(\lambda) = 2\lambda U_n - \lambda^2 V_n + r_n(\lambda),$$

kur

$$|r_n(\lambda)| \le (2/3)\sum_{i=1}^n |(\lambda X_{n,i})^3||h(\lambda X_{n,i})|$$

$$\leq (4/3)|\lambda| D_n \lambda^2 V_n \leq (4/3) c^3 D_n V_n,$$

kas nodrošina $cD_n \leq \frac{1}{2}$.

Ar $T_{n,c}$ un $T^*_{n,c}$ apzīmēsim maksimālās G_n un G^*_n vērtības apgabalā $\Omega_n(c)$, kas ir vienāds ar $\{\lambda : |\lambda| \leq ca_n\}$, un iegūstam

$$|T_{n,c}/a_n - T^*_{n,c}/a_n| \leq (1/a)_n \max\{|r_n(\lambda)| : |\lambda| \leq ca_n\}$$

$$\leq (4/3) c^3 a_n^2 D_n V_n,$$

kamēr $ca_n D_n \leq \frac{1}{2}$. Izvēlēsimies c pietiekami lielu, lai gan $\widehat{\lambda}$, gan λ^* piederētu $\Omega_n(c)$ ar varbūtību lielāku par $1 - \eta$, kādam iepriekš izvēlētam η. Tad

$$P\left\{|\sup G_n/a_n - \sup G^*_n/a_n| \geq \varepsilon\right\} \leq P\left\{(4/3)c^2 a_n^2 D_n V_n \geq \varepsilon\right\}$$

$$+P\{|\widehat{\lambda}| > ca_n\} + P\left\{|\lambda^*| > ca_n\right\} + P\{ca_n D_n > (1/2)\}.$$

Tad lim-sup varbūtību virknei kreisajā pusē ir ierobežota ar 2η. Tā kā η tika izvēlēts patvaļīgi, tad $\sup G_n/a_n$ un $\sup G^*_n/a_n$ ir jābūt vienādiem robežsadalījumiem, kas ir U^2/σ^2. [4] □

Piezīme 9. *Tā kā 4. nodaļā aplūkotie piemēri ir ar dimensiju $p = 1$, tad šeit neapskatīsim Teorēmas 7 pierādījumu vispārējā gadījumā. Tas atrodams materiālā [4].*

4. EL ar novērtētiem parametriem pielietojumi

Šajā nodaļā apskatīsim dažus piemērus, kuros izmantota EL metode ar novērtētiem parametriem. Kā jau iepriekš norādījām, pieņemsim, ka $a_n = 1$ un $U \sim N_p(0, V_1)$, kā arī $m_n(X_i, \theta, h) = m(X_i, \theta, h)/\sqrt{n}$, taču vispārinājums var būt noderīgs citos pielietojumos.

4.1. Integrālis pa blīvuma funkcijas kvadrātu

Nedaudz pieskarsimies rangu testiem, kā piemēram, Vilkoksona rangu testi, jo to efektivitāte lielām izlasēm ir augstāka nekā t-testam vai F-testam. Rangu testi ir robusti, t.i., to statistikas vērtība netiek nozīmīgi ietekmēta ar *outlier* klātbūtni izlasē. Tomēr šo testu asimptotiskais sadalījums, protams, ir atkarīgs no dotās izlases sadalījuma, tāpēc ir nepieciešamība to novērtēt.

Publikācijā [8] J.L. Hodges un E.L. Lehmann pierāda, ka, ja $m/N \to \gamma$, kad $N \to \infty$, kur m - pirmās izlases apjoms un N - novērojumu skaits, tad Hodges-Lehmann lokācijas novērtējums med$(X-Y)$ (mediāna izlases X un Y starpību izlasei) ir asimptotiski normāls ar vidējo vērtību 0 un dispersiju

$$\frac{1}{12\gamma(1-\gamma)\left(\int f^2(x)\mathrm{d}x\right)^2}.$$

Tāpēc šajā piemērā apskatīsim, kā novērtēt parametru $\theta = \int f^2 \mathrm{d}x$, jo Hodges-Lehmann lokācijas novērtējuma dispersija ir proporcionāla $1/\theta^2$ (sk. arī [9] un [10]). Līdzīgi arī Vilkoksona rangu testa pakāpi būtiski ietekmē θ lielums [11].

Tātad $X_1, \ldots, X_n$ *iid* ar nezināmu blīvuma funkciju f_0, kura ir vienmērīgi nepārtraukta, bet ne vienmērīgā sadalījuma. Parametra $\theta_0 = \int f_0^2 \mathrm{d}x$ vērtība bieži tiek pielietota arī dažādās problēmās, kas saistītas ar neparametrisko blīvuma novērtēšanu.

$m(X, \theta, f) = f(X) - \theta$ izvēlamies par novērtējošo funkciju, un kā f_0 *plug-in* novērtējumu izmantosim $\widehat{f}(x) = n^{-1} \sum_{i=1}^{n} k_b(X_i - x)$, kur $k_b(\cdot) = k(\cdot/b)/b$ kodolu blīvuma novērtējums un b joslas platums.

Nosacījumu pārbaude

Vispirms pārbaudīsim, ka novērtējošā funkcija $m(X, \theta, f) = f(X) - \theta$ ir nenovirzīta, t.i.,

$$E(f(X) - \theta) = E(f(X)) - E\theta = \int_{-\infty}^{+\infty} f(x) f(x) \mathrm{d}x - \theta = 0.$$

Lai pārliecinātos, ka ir spēkā nosacījumi $(A0)$ - $(A3)$, definēsim

$$V = \int (f_0 - \theta_0)^2 f_0 \mathrm{dx} = \int f_0^3 \mathrm{d}x - \left(\int f_0^2 \mathrm{d}x \right)^2,$$

kas ir $\sum_{i=1}^n m(X_i, \theta_0, f_0)/\sqrt{n}$ asimptotiskā dispersija, un ir pozitīva, tā kā f_0 nav vienmērīgi sadalīts.

Pārliecināsimies, ka $(A2)$ ir spēkā, kad $V_2 = V$. Tātad

$$n^{-1} \sum_{i=1}^n m^2(X_i, \theta_0, \widehat{f}) = n^{-1} \sum_{i=1}^n \left(\widehat{f}(X_i) - \theta_0 \right)^2 = \int \widehat{f}^2 \mathrm{d}\widehat{F}_n - 2\theta_0 \widehat{\theta} + \theta_0^2,$$

kur $\widehat{F}_n$ empīriskā sadalījuma funkcija un $\widehat{\theta} = n^{-1} \sum_{i=1}^n \widehat{f}(X_i) = \int \hat{f} \mathrm{d}\widehat{F}$. Tad $\int \hat{f} \mathrm{d}\hat{F}$ un $\int \hat{f}^2 \mathrm{d}\widehat{F}$ pēc varbūtības konverģē attiecīgi uz $\int f_0^2 \mathrm{d}x$ un $\int f_0^3 \mathrm{d}x$, kad $b \to 0$ un $nb \to \infty$.

Lai pārbaudītu $(A1)$, nepieciešama rūpīgāka izpēte izteiksmei

$$\widehat{\theta} = n^{-1} \sum_{i=1}^n \widehat{f}(X_i) = n^{-2} \sum_{i,j} k_b(X_i - X_j) = \frac{k(0)}{nb} + \frac{n-1}{n} \widehat{g}.$$

Šeit $\widehat{g} = \widehat{g}(0)$, kur $\widehat{g}(y) = (\frac{n}{2})^{-1} \sum_{i<j} \bar{k}_b(Y_{i,j}, y)$ ir dabiskais kodolu novērtējums blīvuma funkcijai $g(y) = \int f(y+x) f(x) \mathrm{d}x$ no starpības $Y_{i,j} = X_i - X_j$, un $\bar{k}_b(Y_{i,j}, y) = \frac{1}{2}\{k_b(Y_{i,j} - y) + k_b(Y_{i,j} + y)\}$. Hjort ([12]) parādīja, ka $\widehat{g}(y)$ vidējā vērtība ir $g(y) + \frac{1}{2} b^2 g''(y) \int u^2 k(u) \mathrm{d}u + o(b^2)$ un dispersija $\frac{4}{n}\{g^*(y) - g^2(y)\}$ plus bezgalīgi mazas funkcijas ar zemāku kārtu, kur $g^*(y) = \frac{1}{4}\{\bar{g}(y,y) + \bar{g}(y,-y) + \bar{g}(-y,y) + \bar{g}(-y,-y)\}$ un $\bar{g}(y_1, y_2)$ ir kopējā blīvuma funkcija divām saistītām starpībām $(X_2 - X_1, X_3 - X_1)$. No tā seko, ka izteiksmei

$$n^{-1/2} \sum_{i=1}^n m(X_i, \theta_0, \widehat{f}) = \sqrt{n}(\widehat{\theta} - \theta_0)$$

ir vidējā vērtība ar kārtu $O(1/(\sqrt{n}b) + \sqrt{n}b^2)$ un dispersija, kas tiecas uz $4V$. Tas apstiprina $(A1)$ un dod $U \sim N(0, 4V)$, pie nosacījumiem, ka $\sqrt{n}b \to \infty$ un $\sqrt{n}b^2 \to 0$.

Lai pārbaudītu $(A3)$, ievērosim, ka $\widehat{f}(x) \leq b^{-1} k_{\max}$ katram x, kur $k_{\max}$ ir $k(u)$ maksimums. Tātad $\max_{i \leq n} |\widehat{f}(X_i) - \theta_0|$ ir ierobežots ar $b^{-1} k_{\max} + \theta_0$, kas nozīmē, ka $(A3)$ ir spēkā, ja $\sqrt{n}b \to \infty$.

Beidzot, pierādot $(A0)$, mums jāparāda, ka

$$P\{\min_{1 \leq i \leq n} m(X_i, \theta_0, \widehat{f}) < 0 < \max_{1 \leq i \leq n} m(X_i, \theta_0, \widehat{f})\} \to 1.$$

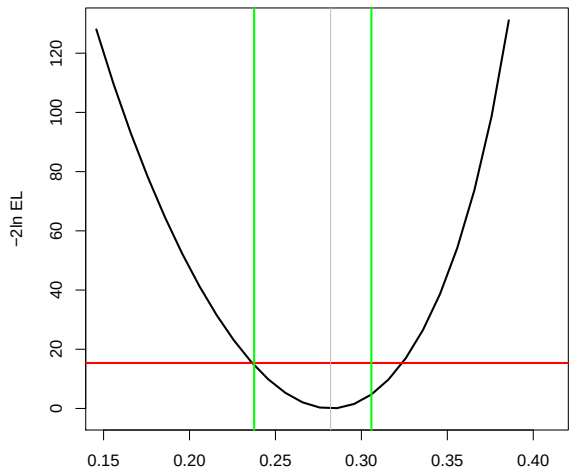
Vispirms apskatīsim

$$\max_{1 \leq i \leq n} m(X_i, \theta_0, \widehat{f}) \geq \max_{1 \leq i \leq n} f_0(X_i) - \max_{1 \leq i \leq n} |\widehat{f}(X_i) - f_0(X_i)| - \theta_0.$$

Ievērosim, ka $\max_{1\le i\le n}|\widehat{f}(X_i)-f_0(X_i)|\to 0$ g.d. tā kā $\widehat{f}$ vienmērīgi nepārtraukta ar piemērotu kodolu, piemēram, standartnormālo blīvuma funkciju, jo pieņēmām, ka f_0 ir vienmērīgi nepārtraukta. Turklāt, $\max_{1\le i\le n} f_0(X_i)\to \sup_t f_0(t)>\theta_0$ g.d., tā kā f_0 nepārtraukta un nav vienmērīgā sadalījuma, tāpēc $P\{\max_{1\le i\le n} m(X_i,\theta_0,\widehat{f})>0\}\to 1$. Līdzīgā veidā varam apskatīt $\min_{1\le i\le n} m(X_i,\theta_0,\widehat{f})$ un secināt, ka $-2\ln \mathrm{EL}_n(\theta_0,\widehat{f})\to 4\chi_1^2$ pēc sadalījuma [4].

Simulācijas

Teorētiski pierādījām, ka EL metode ar novērtētu traucējošo parametru $\widehat{f}$ parametram $\theta_0=\int f_0^2\mathrm{d}x$ strādā. Apskatīsim, kā tas izskatās praktiski. Simulēsim standartnormālā sadalījuma izlasi ar apjomu $n=100$. Tā kā mums ir zināms sadalījums, tad varam aprēķināt īsto θ_0 vērtību, kas ģenerētajai izlasei ir 0.282. 3. attēlā redzama EL metodes statistika parametram θ_0, 95% ticamības intervāls, kas atrasts, izmantojot Teorēmu 7, t.i., $-2\ln \mathrm{EL}_n(\theta_0,\widehat{f})\to_d 4\chi_1^2$, un 95% Butstrapa ticamības intervāls (ar zaļu krāsu). EL ticamības intervāls ir $[0.2364, 0.3238]$, tomēr Butstrapa ticamības intervāls ir šaurāks – $[0.2376, 0.3058]$. Lai arī ticamības intervāls ar EL metodi nav optimālākais šajā gadījumā, EL metodes θ_0 novērtējums $\widehat{\theta}=0.2821$ ir ļoti tuvs īstajam parametram θ_0, tātad metode strādā labi šim piemēram.



3. att. Ticamības intervāli ar EL un Butstrapa metodi parametram $\theta_0=\int f_0^2\mathrm{d}x$

4.2. Atlikumu sadalījumi neparametriskajā regresijā

Pārlūkojot globālo tīmekli, var atrast ļoti daudz materiālus, kas veltīti neparametriskai regresijas atlikumu sadalījuma novērtēšanai. Jāpiemin, ka saistīta, bet specifiskāka problēma ir atlikumu dispersijas novērtēšana homoskedastiskos neparametriskās regresijas modeļos, kas guvusi vēl lielāku autoru uzmanību. Viens no piedāvātajiem novērtējumiem ir empīriskā sadalījuma funkcija novērtētiem atlikumiem, kurus iegūst, izmantojot neparametrisko regresijas funkcijas novērtējumu, kas šajā piemērā ir tā sauktais *nuisance* parametrs.

Aplūkosim modeli $Y = \mu(X) + \varepsilon$, kur X un ε ir neatkarīgi, ε sadalījuma funkcija F_ε nav zināma, un $\mu(\cdot)$ ir nezināma regresijas funkcija. Konstruēsim EL ticamības intervālus parametram $\theta_0 = F_\varepsilon(z) \in (0,1)$ fiksētā punktā z. Tiek ieviesti tādi paši pieņēmumi kā Akritas un Van Keilegom publikācijā [13] - F_ε ir nepārtraukta, $\mu(\cdot)$ ir gluda funkcija, un X ir ierobežota. Vienkāršībai ierobežosim X intervālā $(0,1)$.

Izmantosim Nadaraya-Watson novērtējumu $\widehat{\mu}(x) = \sum_{i=1}^n W_{n,i}(x;b_n)Y_i$ regresijas funkcijas μ novērtēšanai, kur $W_{n,i}(x;b_n) = k_{b,x}(X_i)/\sum_{j=1}^n k_{b,x}(X_j)$, un $k_{b,x}(u) = b^{-1}k((u-x)/b)$. Šajā piemērā izmantosim novērtējošo funkciju $m(X,Y,\theta,\mu) = I_{\{Y-\mu(X)\le z\}} - \theta$, kur I ir indikatorfunkcija.

Nosacījumu pārbaude

Tāpat kā iepriekš pārbaudīsim, vai funkcija $m(X,Y,\theta,\mu) = I_{\{Y-\mu(X)\le z\}} - \theta$ ir nenovirzīta:

$$E(I_{\{Y-\mu(X)\le z\}} - \theta_0) = E(I_{\{Y-\mu(X)\le z\}}) - E\theta_0 = E(I_{\{\varepsilon\le z\}}) - \theta_0 = F_\varepsilon(z) - \theta_0 = 0.$$

Tālāk pārbaudīsim nosacījumus $(A0)$ - $(A3)$, kas nepieciešami, lai varētu pielietot Teorēmu 7 . Pirmkārt $(A1)$ seko no $\widehat{\theta} = n^{-1}\sum_{i=1}^n I_{\{\hat{\varepsilon}_i\le z\}}$, kur $\hat{\varepsilon}_i = Y_i - \widehat{\mu}(X_i)$, asimptotiskās normalitātes, kas aprakstīta [13]: $\sqrt{n}\{\widehat{F}_\varepsilon(z) - F_\varepsilon(z)\} = n^{-1/2}\sum_{i=1}^n m(X_i,Y_i,\theta_0,\widehat{\mu}) \to_d N(0,V_1)$, un V_1 definēta [13].

Nosacījums $(A2)$ ir spēkā ar $V_2 = \theta_0(1-\theta_0)$, ja $0 < \theta_0 < 1$. Arī $(A3)$ izpildās, tā kā funkcija $\sqrt{n}m_n$ ir vienmērīgi ierobežota ar 1. Visbeidzot, $(A0)$ ir tūlītējas sekas no fakta, ka $P\{Y - \widehat{\mu}(X) \le z\}$ konverģē uz $F_\varepsilon(z)$, kas seko no Teilora izvirzījuma un $\widehat{\mu}$ vienmērīgās nepārtrauktības. Tā kā $F_\varepsilon(z)$ ir strikti ierobežots $(0,1)$, tad

$$P\{\exists \quad 1 \le i,j \le n : Y_i - \widehat{\mu}(X_i) \le z \text{ un } Y_j - \widehat{\mu}(X_j) > z\} \to 1,$$

kas apstiprina $(A0)$.

Publikācijā [4] dots pierādījums, no kura seko, ka $100(1-\alpha)\%$ ticamības intervāls parametram $\theta_0 = F_\varepsilon(z)$ ir $\{\theta : -2\ln \mathrm{EL}_n(\theta,\mu) \geq e_{1-\alpha}\}$, kur $e_{1-\alpha}$ ir $100(1-\alpha)$ kvantile sadalījumam

$$\frac{n\left(n^{-1}\sum_{i=1}^{n} I_{\{Y_i-\widehat{\mu}(X_i)\leq z\}} - \widehat{\theta}\right)^2}{\widehat{\theta}(1-\widehat{\theta})}.$$

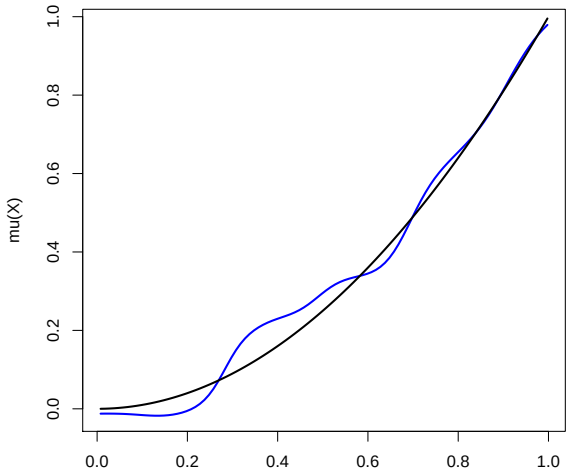
Gludinātā EL metode

Tā kā šim piemēram dotā novērtējošā funkcija $m(X,Y,\theta,\mu) = I_{\{Y-\mu(X)\leq z\}} - \theta$ satur indikatorfunkciju I, tad ir nepieciešams to gludināt. Apskatīsim gludināto *plug-in* EL metodi, kuru lietosim šī piemēra simulācijās. Turklāt Chen un Hall [14] parādīja, ka, pielietojot gludināšanu, var tikt uzlabota ticamības intervālu pārklājumu precizitāte.

Tātad šim piemēram gludinātā novērtējošā funkcija, kuru turpmāk arī izmantosim, ir $m_H(X_i,Y_i,\theta,\mu) = H_b(z-(Y_i-\mu(X_i))) - \theta$, kur $H_b(t) = \int_{u\leq t} K(u)\mathrm{d}u$ un K ir kāda kodola funkcija. Sīkāk ar šo metodi var iepazīties materiālā[14].

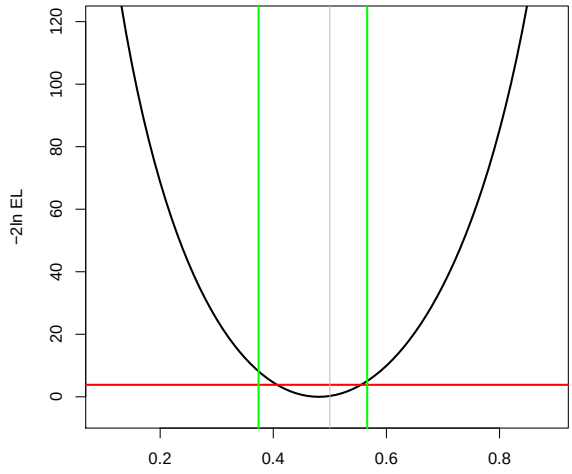
Simulācijas

Pārbaudīsim arī šo piemēru ar simulācijām programmā R. Simulēsim datus kvadrātiskai regresijai, t.i., zināms, ka $\mu(x) = x^2$. 4. attēlā redzams Nadaraya-Watson novērtējums funkcijai μ.



4. att. Nadaraya-Watson novērtējums funkcijai $\mu = x^2$

Pielietosim EL metodi parametra $\theta_0 = F_\varepsilon(0)$ atrašanai, tā kā ir zināms atlikumu sadalījums – $N(0, 0.1)$, tad punktā $z = 0$ īstā parametra θ_0 vērtība ir 0.5. EL metode dod parametra θ_0 novērtējumu $\widehat{\theta} = 0.4805$. 5. attēlā redzams 95% ticamības intervāls ar EL metodi – $[0.4067, 0.5551]$ un 95% Butstrapa ticamības intervāls $[0.3739, 0.566]$. Šajā piemērā ar EL atrastais ticamības intervāls ir šaurāks nekā Butstrapa intervāls, tas nozīmē, ka šjā piemērā EL strādā labāk nekā Butstrapa metode.



5. att. Ticamības intervāli ar EL un Butstrapa metodi parametram $\theta_0 = F_\varepsilon(0)$

4.3. Citi piemēri

Šajā nodaļā apskatīsim dažus pielietojumus EL metodei ar novērtētiem parametriem bez teorētiskā pamatojuma par nosacījumu $(A0)$ - $(A3)$ izpildīšanos, taču ar to var iepazīties publikācijā [4].

Funkcionāļi izdzīvošanas analīzē

Wang un Jing ([5])izstrādāja *plug-in* empīriskās ticamības funkcijas metodes versiju funkcionāļu klasei izdzīvošanas analīzē ar cenzorētu datu klātbūtni. Apzīmēsim riska funkciju un cenzorēto riska funkciju attiecīgi ar F un G. Interesējošais parametrs ir lineārs funkcionālis no F formā $\theta = \theta(F) = \int_0^\infty \xi(t)\mathrm{d}F(t)$, kur $\xi(t)$ ir (zināma) nenegatīva mērojama funkcija un $\theta(F)$ ir galīgs. Šajā piemērā novērtējošā funkcija $m_n = n^{-1/2}m$, un

$$m(Z, \delta, \theta, G) = \frac{\xi(Z)\delta}{1 - G(Z)} - \theta,$$

$Z = \min(X, Y)$, $\delta = I_{\{X<Y\}}$, $Y \sim G$. Šeit tiek pieņemts, ka $X \sim F$ un $Y \sim G$ neatkarīgi. $\widehat{G}_n$ ir Kaplan-Meier novērtējums, kas kalpo par *plug-in* novērtējumu šajā piemērā, savukārt $\widehat{\theta}$ ir formā $\widehat{\theta} = \theta(\widehat{F}_n)$, kur $\widehat{F}_n$ arī ir F Kaplan-Meier novērtējums.

Simetriskas sadalījuma funkcijas

Pieņemsim, ka F ir sadalījuma funkcija gadījuma lielumam X, kas ir simetrisks ap kādu nezināmu lokācijas punktu a, tāpēc $F(x) = 1 - F(2a - x)$ katram x. Apskatīsim parametra $\theta_0 = F(x)$ novērtējumu fiksētā punktā x no n *iid* novērojumiem ar sadalījuma funkciju F. Novērtējošā funkcija satur 2 komponentes (pirmā no tām ir parastā novērtējošā funkcija, bet otrā izmanto simetrijas pieņēmumu): $m_n = n^{-1/2}m$, kur

$$m(X, \theta, a) = \begin{pmatrix} I_{\{X \le x\}} - \theta \\ I_{\{X > 2a - x\}} - \theta \end{pmatrix}.$$

Par a *plug-in* novērtējumu $\widehat{a}$ tiek izmantota izlases mediāna.

Secinājumi

Darbā galvenā uzmanība tika vērsta uz EL metodi ar novērtētiem (*plug-in*) parametriem, taču, lai varētu izprast tās problemātiku, bija nepieciešamība iepazīties ar tradicionālo empīriskās ticamības metodi un galvenajiem rezultātiem, ko izmanto ticamības intervālu konstruēšanā. Ar simulācijām tika pārbaudīta Teorēma 3, un, kā redzams 1. attēlā, EL metodes statistika vidējai vērtībai tiecas uz χ^2_1 sadalījumu. Tas deva pamatu ticamības intervālu konstruēšanas problemātikas izpratnei. Ar ticamības intervālu pārklājumu precizitāti tika pārbaudīts, ka EL ticamības metode strādā līdzīgi kā parametriskie testi, taču būtu vēlams uzlabot pārklājumu precizitāti ticamības intervāliem ar EL metodi nelielu izlašu gadījumā.

Tā kā apskatītie EL metodes pielietojumi ar novērtētiem parametriem ir viendimensionāli, tad publikācijā [4] apskatītie nosacījumi, galvenā teorēma un tās pierādījums tika pārstrādāts no p-dimensionāla gadījuma uz gadījumu, kad $p = 1$. Darbā tika pārbaudīts, vai un kā šie nosacījumi izpildās visvienkāršākajā gadījumā - vidējai vērtībai, kad netiek ieviesti *plug-in* novērtējumi. Tas ļāva secināt, ka dotie nosacījumi vispārina statistikas likumus, kas vienādi un neatkarīgi sadalītiem novērojumiem ar galīgu dispersiju vienmēr ir spēki, piemēram, centrālā robežteorēma vai lielo skaitļu likums.

Plug-in novērtējumu izmantošana var izraisīt sekas, kad pieminētie statistikas likumi var nebūt spēkā, tāpēc 3. nodaļā aprakstīto nosacījumu eksistence ir būtiska. 4. nodaļā apskatīts pierādījums tam, ka šie nosacījumi izpildās abiem sīkāk apskatītajiem piemēriem. Tas ļauj izmantot Teorēmu 7 ticamības intervālu konstruēšanai piemēros apskatītajiem parametriem. Pirmajā piemērā EL ar novērtētiem parametriem ticamības intervāls ir nedaudz plašāks nekā Butstrapa ticamības intervāls, taču EL metodes parametra novērtējums ir ļoti precīzs. Savukārt otrajā piemērā situācija ir pretēja - EL ticamības intervāls ir šaurāks nekā Butstrapa intervāls, kas, iespējams, ir gludinātās EL pielietošanas rezultāts, jo kā pieminēts, tādā veidā var tikt uzlabota ticamības intervālu pārklājumu precizitāte.

Tā kā empīriskās ticamības metode tiek plaši pielietota tikai nesen, tad vēl ir daudz iespējamu metodes uzlabojumu, kurus pielietojot, šī metode būtu vēl spēcīgāks konkurents parametriskajām metodēm.

Izmantotā literatūra un avoti

[1] A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120, 1990.

[2] A.B. Owen. *Empirical likelihood*. CRC press, 2001.

[3] D.R. Thomas and G.L. Grunkemeier. Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, pages 865–871, 1975.

[4] N.L. Hjort, I.W. McKeague, and I. Van Keilegom. Extending the scope of empirical likelihood. *Ann. Statist*, 37(3):1079–1111, 2009.

[5] Q.H. Wang and B.Y. Jing. Empirical likelihood for a class of functionals of survival distribution with censored data. *Annals of the Institute of Statistical Mathematics*, 53(3):517–527, 2001.

[6] M. Dekking, C. Kraaikamp, and HP Lopuhaa. *A Modern Introduction to Probability and Statistics: Understanding why and how*. Springer Verlag, 2005.

[7] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325, 1994.

[8] JL Hodges Jr and E.L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, 1963.

[9] E.L. Lehmann and G. Casella. *Theory of point estimation*. Springer Verlag, 1998.

[10] N. Inagaki. The asymptotic representation of the Hodges-Lehmann estimator based on Wilcoxon two-sample statistic. *Annals of the Institute of Statistical Mathematics*, 25(1):457–466, 1973.

[11] E.L. Lehmann and HJM D'abrera. *Nonparametrics: statistical methods based on ranks*, volume 204. Holden-Day San Francisco, 1975.

[12] N.L. Hjort and U. i Oslo. *Towards semiparametric bandwidth selectors for kernel density estimators*. Department of Mathematics, University of Oslo, 1999.

[13] M.G. Akritas and I. Van Keilegom. Non-parametric Estimation of the Residual Distribution. *Scandinavian Journal of Statistics*, 28(3):549–567, 2001.

[14] S.X. Chen and P. Hall. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, 21(3):1166–1181, 1993.

1. Pielikums

1.1. Programmas R kods Owena teorēmas pārbaudei

```
R_FF<-c()
n<-100
mu<-0
for (k in 1:1000)
{
izl<-rnorm(n,0,1)

izl_sort<-c()
izl_sort<-sort(izl)

lambda_l<-(1-1/n)/(mu-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(mu-izl_sort[1])  #Lambda augseja robeza

f.lam<-function(lambda)
{
sum((izl-mu)/(1+lambda*(izl-mu)))
}
f.lam2<-Vectorize(f.lam)

lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(izl-mu)))
R_FF[k]<--2*log(prod(n*p))
}

hist(R_FF,prob=TRUE,main="",xlab="-2ln EL",col="blue")
xx<-seq(0,12,by=0.01)
lines(xx,dchisq(xx,1),col="red",lwd=2)
```

1.2. Programmas R kods ticamības intervāliem vidējai vērtībai

```
n<-100
izl<-rnorm(n,0,1)

R_FF<-function(mu)
{
izl_sort<-c()
izl_sort<-sort(izl)
lambda_l<-(1-1/n)/(mu-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(mu-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
sum((izl-mu)/(1+lambda*(izl-mu)))
}
f.lam2<-Vectorize(f.lam)
#lam<-seq(lambda_l,lambda_u,by=0.01)
#plot(lam,f.lam2(lam),type="l")
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(izl-mu)))
R_FF<--2*log(prod(n*p))
R_FF
}
R_FF2<-Vectorize(R_FF)
r.ff<-seq(min(izl)+1.33,max(izl)-1.57,by=0.01)
plot(r.ff,R_FF2(r.ff),type="l",lwd=2,ylab="-2ln EL",xlab="",main="")
xx<-function(virkne)
{
qchisq(0.95,1)
}
xx2<-Vectorize(xx)
lines(r.ff,xx2(r.ff),lwd=2,col="red")
```

```
R_FF3<-function(mu)
{
R_FF3<-R_FF(mu)-qchisq(0.95,1)
}
R_FF4<-Vectorize(R_FF3)
lines(r.ff,R_FF4(r.ff))
xx3<-function(virkne)
{
0
}
xx4<-Vectorize(xx3)
lines(r.ff,xx4(r.ff))
apak.rob<-uniroot(R_FF3,c(min(izl)+0.1,
optimize(R_FF,c(min(izl),max(izl)))$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(izl),max(izl)))
$minimum,max(izl)-0.1))$root
apak.rob
aug.rob
```

1.3. Programmas R kods pārklājumu precizitātei ticamības intervāliem

```
n<-100
prec<-function(izl,mu_0)
{
R_FF<-function(mu)
{
izl_sort<-c()
izl_sort<-sort(izl)
lambda_l<-(1-1/n)/(mu-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(mu-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
```

```
sum((izl-mu)/(1+lambda*(izl-mu)))
}
f.lam2<-Vectorize(f.lam)
#lam<-seq(lambda_l,lambda_u,by=0.01)
#plot(lam,f.lam2(lam),type="l")
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(izl-mu)))
R_FF<--2*log(prod(n*p))
R_FF
}
R_FF3<-function(mu)
{
R_FF3<-R_FF(mu)-qchisq(0.95,1)
}
apak.rob<-uniroot(R_FF3,c(min(izl)+0.1,
optimize(R_FF,c(min(izl),max(izl)))$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(izl),
max(izl)))$minimum,max(izl)-0.1))$root
if ((mu_0>apak.rob)&(mu_0<aug.rob)){prec<-1} else {prec<-0}
}
prec1<-replicate(10000,prec(rnorm(n,0,1),0))
parkl.prec<-sum(prec1)/10000
parkl.prec
k<-function(dati)
{
k1<-t.test(dati)$conf.int
b<-k1[1]*k1[2]
if (b<0) 1 else 0
}
prec.t<-replicate(10000,k(rnorm(n,0,1)))
sum(prec.t)/10000
```

1.4. Programmas R kods integrālim pa blīvuma funkcijas kvadrātu

```
library(sm)
n<-100
izl<-rnorm(n,0,1)
b<-hcv(izl) #gludinosais parametrs
###Atrast isto theta parametru
d<-function(x) dnorm(x)^2
theta0<-integrate(d,-5,5)
theta0
plot(X,Y)
help(plot)
###Butstrapa ticamibas intervali
theta1<-function(dati)
{
f00<-function(x) #kodola blivuma f-jas novertejums
{
1/n/b*sum(dnorm((dati-x)/b))
}
f111<-Vectorize(f00)
d1<-function(x) f111(x)^2
integrate(d1,-5,5)$value
}
B<-1000
theta.boot<-replicate(B,theta1(sample(izl,replace=TRUE)))
se.boot<-var(theta.boot) #S^2
apak.rob.boot<-theta1(izl)-qnorm(0.95)*sqrt(se.boot)
aug.rob.boot<-theta1(izl)+qnorm(0.95)*sqrt(se.boot)
apak.rob.boot
aug.rob.boot

####Intervaali
```

```
R_FF<-function(theta)
{
f0<-function(x) #kodola blivuma f-jas novertejums
{
1/n/b*sum(dnorm((izl-x)/b))
}
f1<-Vectorize(f0)
izl_sort<-c()
izl_sort<-sort(f1(izl))
lambda_l<-(1-1/n)/(theta-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(theta-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
sum((f1(izl)-theta)/(1+lambda*(f1(izl)-theta)))
}
f.lam2<-Vectorize(f.lam)
#lam<-seq(lambda_l,lambda_u,by=0.1)
#plot(lam,f.lam2(lam),type="l")
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(f1(izl)-theta)))
-2*log(prod(n*p))
}
R_FF2<-Vectorize(R_FF)
r.ff<-seq(min(f1(izl))+0.099,max(f1(izl))-0.037,by=0.01)
plot(r.ff,R_FF2(r.ff),type="l",ylab="-2ln EL",xlab="",
lwd=2,ylim=c(-2,130),xlim=c(0.15,0.41))
abline(h=4*qchisq(0.95,1),lwd=2,col="red")
R_FF3<-function(theta)
{
R_FF3<-R_FF(theta)-4*qchisq(0.95,1)
}
R_FF4<-Vectorize(R_FF3)
```

```
lines(r.ff,R_FF4(r.ff),col="blue")
abline(h=0)
apak.rob<-uniroot(R_FF3,c(min(f1(izl))+0.1,optimize(R_FF,c(min(f1(izl)),
max(f1(izl))))$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(f1(izl)),max(f1(izl))))
$minimum,max(f1(izl))-0.1))$root
apak.rob
aug.rob
abline(v=0.28209,col="gray")
abline(v=apak.rob.boot,col="green",lwd=2)
abline(v=aug.rob.boot,col="green",lwd=2)

EL<-optimize(R_FF,c(min(f1(izl)),max(f1(izl))))$minimum
EL #Neparametriskas ticamibas metodes novertejums
```

1.5. Programmas R kods regresijas atlikumu sadalījumiem

```
library(sm)
n<-100
X<-runif(n,0,1)
eps<-rnorm(n,0,0.1)
Y<-X^2+eps
scatterplot(X,Y)
pnorm(0,0,0.1)
###mu noveerteetais
b1<-hcv(X) #gludinosais parametrs
mu<-function(x)
{
sum((1/b1*dnorm((X-x)/b1))/sum(1/b1*dnorm((X-x)/b1))*Y)
}
mu1<-Vectorize(mu)
m<-seq(min(X),max(X),by=0.01)
plot(m,mu1(m),type="l",lwd=2,col="blue",ylab="mu(X)",xlab="")
```

```
m1<-function(x) x^2
points(m,m1(m),type="l",lwd=2)
eps.nov<-Y-mu1(X)
z<-0
b2<-hcv(eps.nov)
H<-pnorm((z-eps.nov)/b2) #Sadaliijuma f-jas noveerteejums

izl_sort<-c()
izl_sort<-sort(H)

R_FF<-function(theta)
{
lambda_l<-(1-1/n)/(theta-izl_sort[n])  #Lambda apakseja robeza
lambda_u<-(1-1/n)/(theta-izl_sort[1])  #Lambda augseja robeza
f.lam<-function(lambda)
{
sum((H-theta)/(1+lambda*(H-theta)))
}
f.lam2<-Vectorize(f.lam)
#lam<-seq(lambda_l,lambda_u,by=0.1)
#plot(lam,f.lam2(lam),type="l")
lambda<-uniroot(f.lam2,c(lambda_l,lambda_u))$root
p<-1/(n*(1+lambda*(H-theta)))
-2*log(prod(n*p))
}
R_FF2<-Vectorize(R_FF)
r.ff<-seq(min(H)+0.1,max(H)-0.1,by=0.01)
plot(r.ff,R_FF2(r.ff),type="l",lwd=2,main="",
ylab="-2ln EL",xlab="",ylim=c(-5,120))

abline(h=qchisq(0.95,1),col="red",lwd=2)
R_FF3<-function(theta)
```

```
{
R_FF3<-R_FF(theta)-qchisq(0.95,1)
}
R_FF4<-Vectorize(R_FF3)
lines(r.ff,R_FF4(r.ff),col="blue")
abline(h=0)
apak.rob<-uniroot(R_FF3,c(min(H)+0.1,optimize(R_FF,c(min(H),max(H)))
$minimum))$root
aug.rob<-uniroot(R_FF3,c(optimize(R_FF,c(min(H),max(H)))
$minimum,max(H)-0.1))$root
apak.rob
aug.rob
theta_nov<-optimize(R_FF,c(min(H),max(H)))$minimum
theta_nov

###Butstrapa ticamibas intervali
z<-0
w<-0
mu<-function(t,v)
{
k<-function(x)sum((1/b1*dnorm((t-x)/b1))
/sum(1/b1*dnorm((t-x)/b1))*v)
k1<-Vectorize(k)
eps.nov.boot<-Y-k1(X)
for (i in 1:n)
if (eps.nov.boot[i]<=0) w<-w+1
w/n
}
B<-1000
theta.boot<-replicate(B,mu(sample(X,replace=TRUE),
sample(Y,replace=TRUE)))
se.boot<-var(theta.boot) #S^2
```

```
apak.rob.boot<-mu(X,Y)-qnorm(0.95)*sqrt(se.boot)
aug.rob.boot<-mu(X,Y)+qnorm(0.95)*sqrt(se.boot)
apak.rob.boot
aug.rob.boot
abline(v=pnorm(0),col="gray")
abline(v=apak.rob.boot,col="green",lwd=2)
abline(v=aug.rob.boot,col="green",lwd=2)
```

Kursa darbs "Empīriskā ticamības funkcija ar novērtētiem parametriem" izstrādāts LU Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Leonora Pahirko

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts kursa gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)