LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

NEPARAMETRISKĀS REGRESIJAS KĻŪDAS DISPERSIJAS NOVĒRTĒŠANA

KURSA DARBS

Autors: **Maksims Korotejevs**

Stud. apl. mk08265

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2012. g. 27. janvārī

Saturs

Apzīmejumi

$\widehat{\sigma}_{RICE,n}(x)$ lokālais klasiskais Rice novērtējums

$\widehat{\sigma}_{MSD,n}(x)$ lokālais M - novērtējums ar MSD funkciju

$\widehat{\sigma}_{MBT,n}(x)$ lokālais M - novērtējums ar BT funkciju

$\widehat{\sigma}_{n}(x)$ novērtējums no Larry Wasserman grāmatas

...

Ievads

Nodarbojoties ar neparametrisko regresiju svarīgi ir novērtēt kļūdu dispersiju. To var izmantot konstruējot ticamības intervālus dispersijai. Dispersijas kļūdu var novērtēt kā konstanti vai kā funkciju. Vispārīgi regresijai ir svarīgs nosacījums par homoskedastitāti. Sarežģītākais gadījums ir heteroskedastiska modeļa gadījums, kad dispersija nav konstanta. Larry Wasserman savā grāmatā [1] aprakstīja dispersijas novērtējumu, kuru izgudroja Yu un Jones [2], dispersijas novērtējumam pielietojot divas reizes neparametrisko regresiju. Savukārt, Brown un Levin [3] vispārināja Rice [4] dispersijas novērtējumu no homoskedastiska modeļa līdz heteroskedastiskam modelim izmantojot kodola svarus. Samērā nesen tika piedāvāti lokālie M robusti, tā saucamie M - novērtējumi. Kad datiem ir izlecēji, nepieciešams izmantot robusta statistiku kļūdas dispersijas funkcijas novērtējumam, lai varētu novērtēt regresijas funkciju šiem datiem. Mēs apskatīsim lokālos M - novērtējumus dispersijai. Lai novērstu izlēcēju efektu uz neparametrisko modeli, izmantosim Rice [4] piedāvāto izteiksmi un pārveidoto izteiksmi ar robusta statistiku, kuru piedāvā Boente, Fraiman un Meloche [5].

Robusta statistikas pielietojumam ir liela nozīmē šajā darbā, tāpēc apskatīsim to derīgumu sīkāk. Dispersijai lietojam robusta statistiku, lai ievērotu izlēcējus (Hanning un Lee) [6], lai datiem būtu iespēja izmantot regresijas funkciju(Hardle un Gasser [7], Hardle un Tsybakov [8], Boente un Fraiman) [9], lai uzlabotu precizitāti izvēloties atbilstošu h garumu, kad novērtējam regresijas funkciju r (Boente [5], Cantoni un Ronchetti [10], Leung [11]).

Tātad robusta statistikas novērtējums, tiek plaši izmantots dispersijas funkcijas novērtēšanai neparametriskā regresijas modelī ar izlēcējiem. Šāds novērtējums ir ļoti noderīgs rēķinot robusta M - novērtējumus regresijas funkcijai (to apskatīja Hardle un Gasser [7], Hardle un Tsybakov [8], Boente un Fraiman [9]).

Šī darba mērķi ir konstruēt vecus un jaunus novērtējumus programmā R, veidojot kļūdas līdzīgi kā publikācijā uz piesārņotiem modeļiem, salīdzināt vairākus novērtējumus, reāliem datiem pielietot augstāk minētos novērtējumus.

Darbs satur trīs nodaļas. Pirmajā nodaļā tiek aplūkota neparametriska un parametriska lineāra regresija. Otrā nodaļa veltīta lokāliem M - novērtējumiem. Trešā nodaļa satur praktisku metodes pielietojumu.

1. Neparametriska un parametriska lineāra regresija

1.1. Parametriska lineāra regresija

Lai problēmu labāk saprastu apskatīsim parametrisko lineāro regresiju. Pieņemsim, ka mums ir dati $(x_1, Y_1), ..., (x_n, Y_n))$, kur $Y_i \in \mathbb{R}$ un $x_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$. Lineāras regresijas modelis ir:

$$Y_i = r(x_i) + \varepsilon_i \equiv \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, i = 1, ..., n,$$

kur $E(\varepsilon_i) = 0$ un $D(\varepsilon_i) = \sigma^2$. Būtiski ir novērtēt parametrus β_j. Lai to izdarītu ieviesīsim matricu:

$$\begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix},$$

kur $x_{i1} = 1$.

Pārrakstīsim Y, ε un β vektor formā: $Y = (Y_1, ..., Y_n)^T$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ un $\beta = (\beta_1, ..., \beta_p)^T$. Tad iegūst

$$Y = X\beta + \varepsilon.$$

Izmantosim mazāko kvadrātu metodi, lai iegūstu $\widehat{\beta} = (\widehat{\beta_1}, ..., \widehat{\beta_p})^T$, kurš minimizē RSS (atlikumu kvadrāt summa):

$$RSS = (Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2.$$

Tā kā $X^T X$ ir ar pilnu rangu, t.i., matrica $X^T X$ nav deģenerēta $\Rightarrow \exists \Big(X^T X \Big)$ inversa matrica. Līdz ar to

$$\widehat{\beta} = \Big(X^T X \Big)^{-1} X^T Y.$$

$r(x)$ novērtējums pie $x = (x_1, ..., x_p)^T$ ir

$$\widehat{r}_n(x) = \sum_{j=1}^{p} \widehat{\beta}_j x_j = x^T \widehat{\beta}.$$

No tā seko, ka $(\widehat{r}_n(x_1), ..., \widehat{r}_n(x_n))^T$ varētu būt pierakstīts, šādi

$$r = X\widehat{\beta} = LY,$$

kur

$$L = X\Big(X^T X\Big)^{-1} X^T.$$

Savukārt izteiksmi $\widehat{\varepsilon} = Y - r$ sauc par kļūdu vektoru, matrica L ir simetriska $L = L^T$ un idempotenta $L^2 = L$. Parametrus skaits p tiek noteikts no matricas L pēdas, kas ir diagonāl elementu summa pēc izteiksmes:

$$p = tr(L).$$

Jebkuram $x = (x_1, ..., x_p)^T$ varam rakstīt

$$\widehat{r}_n(x) = \ell(x)^T Y = \sum_{i=1}^{n} \ell_i(x) Y_i,$$

kur

$$\ell_i(x)^T = x^T\Big(X^T X\Big)^{-1} X^T.$$

σ^2 nenovirzīts novērtējums ir

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \widehat{r}_n(x_i))^2}{n-p}.$$

1.2. Neparametriska regresija

Doti n novērojumu pāri $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$. Skaidrojošais mainīgais Y ir atkarīgs no x pēc vienādojuma

$$Y_i = r(x_i) + \varepsilon_i, E\varepsilon_i = 0, i = 1, ..., n, \tag{1.1}$$

kur $r(x)$ ir regresijas funkcija. Regresijas funkcijas $r(x)$ novērtējumu apzīmēsim ar $\widehat{r}_n(x)$. Pieņemsim, ka dispersija $D(\varepsilon_i) = \sigma^2$ ir atkarīga no x. Pēc modeļa (1.1) uzskatam, ka x_i ir fiksēti. Savukārt mēs varam aplūkot novērojumus $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ kā gadījuma lielumu šāda gadījumā mēs rakstīsim datus, kā $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ un $r(x)$ tiek interpretēts, kā nosacīta matemātiska cerība:

$$r(x) = E(Y|X = x).$$

Definīcija 1. Novērtētājs $\widehat{r}_n$ ir r lineārais gludinātais, ja $\forall x\ \exists l(x) : l(x) = (l_1(x), ..., l_n(x))^T$ tāds, ka

$$\widehat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i.$$

Piezīme: $\forall x : \sum_{i=1}^{n} l_i(x) = 1$.

Definīcija 2. Pieņemsim, ka $h > 0$, kuru sauc par gludinošo parametru. Nadaraya - Watsona kodola novērtētājs ir definēts sekojoši

$$\widehat{r}_n(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

kur K ir kodols un $w_i(x)$ ir svari

$$w_i(x) = \frac{K\Big(\frac{x - x_i}{h}\Big)}{\sum_{j=1}^{n} K\Big(\frac{x - x_j}{h}\Big)}, \tag{1.2}$$

šeit $K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. Gludinošā parametra h izvēle ir svarīga problēma. Eksistē daudz metodes kā to var izdarīt, piemēram, cross-validation, rules of thumb, bootstrap un plug-in metodes.

Ideālā gadījumā parametru h izvēlas tāda veidā, lai minimizētu izteiksmi

$$R(h) = E\Big(\frac{1}{n}\sum_{i=1}^{n} \widehat{r}_n(x_i) - r(x_i)\Big)^2.$$

Bet $R(h)$ ir atkarīgs no nezināmas funkcijas $r(x)$. Tāpēc ņemsim $\widehat{R}(h)$, tādu lai minimizētu:

$$\widehat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \widehat{r}_n(x_i))^2.$$

Savukārt izmantojot plug-in metodi gludinošo parametru h novērtē šādi

$$\widehat{h} = \Big(\frac{R(h)}{\mu_2(K)^2 \widehat{\psi}_4(r) n}\Big)^{\frac{1}{5}},$$

kur $\mu_2(h) = \int z^2(h) dh$ un $\widehat{\psi}_4(r)$ ir kodola novērtējums. Pēc būtības plug-in metode balstās uz to, ka izteiksmē nezināmo parametru, aizvieto ar novērtējumu.

1.3. Dispersijas novērtējums

Apskatīsim aprakstītu Larry Wasserman [1] grāmatas standart metodi σ^2 novērtējumam, kuru izgudroja Yu un Jones [2]. Pieņem, ka

$$Y_i = r(x_i) + \sigma(x_i)\varepsilon_i.$$

Apzīmēsim, ka $Z_i = \log(Y_i - r(x_i))^2$ un $\delta_i = \log(\varepsilon_i^2)$.

Tādējādi dispersijas novērtēšanas procedūra ir sekojoša:

a) Novērtē regresijas funkciju $r(x)$ ar Nadaraya-Watsona neparametrisko metodi, lai iegūtu novērtējumu $\widehat{r}_n(x)$.

b) Aprēķinam $Z_i = \log(Y_i - \widehat{r}_n(x_i))^2$.

c) Veic regresiju Z_i pret x_i, iegūst $\log(\sigma^2(x))$ novērtējumu $\widehat{\sigma}^2$

$$Y_i - r(x_i) = \sigma(x_i)\varepsilon_i,$$

$$(Y_i - r(x_i))^2 = \sigma^2(x_i)\varepsilon_i^2,$$

$$\log(Y_i - r(x_i))^2 = \log(\sigma^2(x_i)\varepsilon_i^2),$$

$$\log(Y_i - r(x_i))^2 = \log(\sigma^2(x_i)) + log(\varepsilon_i^2),$$

tā, kā $Z_i = \log(Y_i - r(x_i))^2$ un $\delta_i = \log(\varepsilon_i^2)$

$$=> Z_i - \delta_i = \log(\sigma^2(x_i)),$$

$$\widehat{\sigma}^2(x_i) = e^{Z_i - \delta_i},$$

2. Dispersijas lokālie M - novērtējumu

Lai labāk saprastu, kas ir M - novērtējumi aplūkosim populārus Hubera novērtējumus un lokālos M novērtējumus.

2.1. Hubera novērtējumi

Pieņemsim, ka novērojumi x_i ir atkarīgi no "īstas vērtības" un gadījuma kļūdas u_i, μ ir nezināmais parametrs. Pieņemsim, ka kļūdas ir aditīvas, t.i,

$$x_i = \mu + u_i, i = 1, ..., n,$$

kur $u_1, ..., u_n$ ir neatkarīgi gadījuma lielumi ar sadalījumu funkciju F_0. Tad šādu konstrukciju sauksim par lokālo modeli. Līdz ar to $x_1, ..., x_n$ ir neatkarīgi ar sadalījuma funkciju

$$F(x) = F_0(x - \mu)$$

un mēs sākam, ka x_i ir neatkarīgi identiski sadalīti gadījuma lielumi.

M - novērtējumi ir plāša statistisko novērtējumu klase. M - novērtējumi, piemēram, ir mazāko kvadrātu novērtējumi. 1964. gadā Peter Huber [12] piedāvāja minimizēt izteiksmi

$$\sum_{i=1}^{n} \rho(x_i, \mu),$$

kur funkciju ρ izvēlas tāda veidā, lai datiem ņemtiem no zināma sadalījuma nodrošinātu vēlamas novērtējumu īpašības (nenovirzītibu un būtiskumu) un pietiekoši lielu izturību pret izlēcējiem.

Pieņemsim, ka F_0 ir u_i sadalījuma funkcija ar blīvuma funkciju $f_0 = F_0'$. Kopēja blīvuma funkcija (likehood funkcija) novērojumiem ir

$$L(x_1, ..., x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu).$$

Šīs problēmas atrisinājums ir $\widehat{\mu}(x_1, ..., x_n)$ un tas izskatās šādi

$$\widehat{\mu} = \widehat{\mu}(x_1, ..., x_n) = arg \max_{\mu} L(x_1, ..., x_n; \mu).$$

Definīcija 3. Par M novērtējumu sauksim

$$\widehat{\mu} = arg \min_{\mu} \sum_{i=1}^{n} \rho(x_i - \mu), \tag{2.1}$$

kur

$$\rho = -\log f_0.$$

Piemēram, ja $F_0 = N(0,1)$, tad

$$f_0(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

kur $\rho(x) = \frac{x^2}{2}$. Līdz ar to (2.1) ir vienāds ar

$$\widehat{\mu} = arg \min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Ja ρ ir diferencējams, tad (2.1) var parakstīt

$$\sum_{i=1}^{n} \psi(x_i - \widehat{\mu}) = 0,$$

kur $\psi = \rho'$. Ja $\rho(x) = \frac{x^2}{2}$, tad $\psi(x) = x$ un (2.1) kļūst

$$\sum_{i=1}^{n} (x_i - \widehat{\mu}) = 0,$$

kur $\widehat{\mu} = \overline{x}$ ir atrisinājums.

Pieņemsim, ka F_0 ir dubults eksponenciālais sadalījums, tad

$$f_0(x) = \frac{1}{2} e^{-|x|},$$

kur $\rho(x) = |x|$ un izteiksme (2.1) ekvivalenta ar

$$\widehat{\mu} = arg \min_{\mu} \sum_{i=1}^{n} |x_i - \mu|.$$

Ja $\rho(x) = |x|$, tad jebkura mediāna no x būs atrisinājums. Pie tam $\psi(x) = sgn(x)$

$$sgn(x) = \begin{cases} -1, \text{ ja } x < 0 \\ 0, \text{ ja } x = 0 \\ 1, \text{ ja } x > 0. \end{cases}$$

Hubers [12] ieviesa $\psi(x,t) = \psi_0(x-t)$, kur

$$\psi_0(z) = \begin{cases} c, \text{ ja } z \geq c \\ z, \text{ ja } |z| < c \\ -c \text{ ja } z \leq -c. \end{cases}$$

Ar tādu nolūku, ka neierobežotām ψ funkcijām piemīt nevēlamas īpašības (izlēcēju ietekme) un ja $c \to \infty$ iegūst vidējo vērtību, bet ja $c \to 0$, iegūst mediānu.

2.2. Dispersijas lokālie M - novērtējumi

Pievērsīsim uzmanību jauniem dispersijas M - novērtējumiem $\widehat{\sigma}_{RICE,n}(x)$, $\widehat{\sigma}_{MSD,n}(x)$ un $\widehat{\sigma}_{MBT,n}(x)$. Lokālo M - novērtējumu metodes būtība balstās uz rezultātu, no homoskedastiskās neparametriskās regresijas modeļa, kuru ieguva Hall[13]. Viņš dispersiju novērtēja šādā veidā:

$$\widehat{\sigma}^2_{r,n} = \frac{1}{n-r} \sum_{i=m_1+1}^{n-m_2} \Big(\sum_{k=-m_1}^{m_2} d_k Y_{i+k} \Big)^2,$$

kur Y_{i+k} jābūt sakārtotam. $\{d_i\}_{i=-m_1}^{m_2}$ ir virkne veidota no starpībām uz reāliem skaitļiem, kura apmierina $\sum_{i=-m_1}^{m_2} d_i = 0$ un $\sum_{i=-m_1}^{m_2} d_i^2 = 1$ ar $d_{-m_1} \neq 0$, $d_{-m_2} \neq 0$ priekš $m_1, m_2 \in Z^+$. Pie tam $r = m_1 + m_2$. Kad $r = 1$, $\widehat{\sigma}^2_{r,n} = \widehat{\sigma}^2_{Rice,n}$, kas ir labi zināms, kā novērtējums pēc Rice[4]:

$$\widehat{\sigma}^2_{Rice,n} = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2.$$

Šīs dispersijas funkciju klases paplašināja līdz heteroskedastiskas neparametriskas regresijas modeļiem 2007. gadā Brown un Levin [3], kuri apskatīja lokālo novērtējumu bāzētu uz kodola svariem.

Vispār, homoskedastiskiem neparametriskiem modeļiem, Ghement[14] novērtēja $\widehat{\sigma}^2_{Rice,n}$ izmantojot M - novērtējumu, definētu, kā atrisinājumu $\widehat{\sigma}_0$ izteiksmei

$$\frac{1}{n-1} \sum_{i=1}^{n-1} \chi\Big(\frac{Y_{i+1} - Y_i}{a\widehat{\sigma}_0} \Big) = b,$$

kur χ ir score funkcija, a pozitīva konstante, b pozitīva konstante, kura dod robust līmeni novērtējumam. Apskatīsim sekojošo neparametriskas regresijas modeli:

$$Y_i = g(x_i) + U_i \sigma(x_i), 1 \leq i \leq n, \tag{2.2}$$

kur $x_0 \leq x_1 \leq ... \leq x_n \leq 1$, σ nezināma dispersijas funkcija, g nezināma regresijas funkcija, kļūdas ir $U_i \sim F_0$ un $U_1, U_2, ..., U_n$ ir iid.

Mūsu mērķis ir novērtēt σ kā funkciju no x. Izrādījās, ka problēma novērtēt funkciju σ heteroskedastiskiem modeļiem ir tik pat svarīga, kā regresijas funkciju g izvēle. Heteroskedastitāte nozīme, ka $\forall i \ D(U_i) \neq const$.

Reprezentēsim funkciju no dispersijas $\sigma(x)$ caur robust statistiku. Šo $\sigma(x)$ novērtējumu sauksim par *lokālos M - novērtējumus no starpībām*. Aplūkosim datus, kuri apmierina 2.2 modeļa nosacījumus. Sākumā ģenerēsim datus ar daudziem izlecēj punktiem. Lai to izdarītu apskatīsim sekojošu apakškopu:

$$\mathcal{P}_\varepsilon(F_0) = \{G | G(y) = (1-\varepsilon)F_0(y) + \varepsilon H(y); H \in D, y \in \mathbb{R}\},$$

kur D ir visu sadalījuma funkcijas kopa, F_0 ir normālais sadalījums, H ir jebkura patvaļīga sadalījuma funkcija, kurā modelē piesārņojumus un $\varepsilon \in [0, 1/2)$.

Vispārīgā veidā definēsim $x \in (0,1)$ dispersijas funkcijai $\sigma(x)$ lokālos M - novērtējumus, no starpībām:

$$\widehat{\sigma}_{M,n}(x) = \inf\{s > 0 | \sum_{i=1}^{n-1} w_{n,i}(x)\chi\Big(\frac{Y_{i+1} - Y_i}{as}\Big) \leq b\}, \tag{2.3}$$

kur $w_{n,i}(x)_{i=1}^{n-1}$ ir svaru funkcijas virkne(piemēram, kodols), χ ir score funkcija, $a \in (0, \infty)$ un $b \in (0,1)$ pie tam

$$E(\chi(Z_1)) = b \text{ un } E\Big(\chi\frac{Z_2 - Z_1}{a}\Big) = b,$$

kur $\{Z_i\}_{i=1,2}$ ir i.i.d. ar sadalījumu $Z_1 \sim F_0$. Parasti $\chi : \mathbb{R} \to \mathbb{R}$.

Izteiksmē (2.3) nepieciešams aprēķināt infinimumu tikai, ja funkcija χ ir pārtraukta, bet ja χ ir nepārtraukta, tad viegli redzēt, ka $\widehat{\sigma}_{M,n}(x)$ apmierina izteiksmi

$$\sum_{i=1}^{n-1} w_{n,i}(x)\chi\Big(\frac{Y_{i+1} - Y_i}{a\widehat{\sigma}_{M,n}(x)}\Big) = b.$$

Aplūkosim trīs lokālos M - novērtējumu piemērus, pēc kuriem tālāk konstruēsim $\sigma(x)$ novērtējumu:

1) Pieņemsim, ka $\chi(x) = x^2, a = \sqrt{2}, b = 1$, tad mēs iegūsim klasisko *local Rice* novērtējumu

$$\widehat{\sigma}_{RICE,n}(x) = \sqrt{\sum_{i=1}^{n-1} w_{n,i}(x)\Big(\frac{Y_{i+1} - Y_i}{\sqrt{2}}\Big)^2}.$$

2) Pieņemsim, ka $\chi(y) = I_{u:|u|>\Phi^{-1}(3/4)}(y), a = \sqrt{2}, b = 1/2$, tad $\widehat{\sigma}_{MSD,n}(x)$ apmierina vienādojumu

$$\sum_{i=1}^{n-1} w_{n,i}(x)\chi\Big(\frac{Y_{i+1} - Y_i}{a\widehat{\sigma}_{MSD,n}(x)}\Big) - b = 0,$$

kuras saknes, būs $\widehat{\sigma}_{MSD,n}(x)$ vērtības.

3) Pieņemsim, ka $\forall c > 0$ izpildās

$$\chi_c(y) = \begin{cases} \{3\Big(\frac{y}{c}\Big)^2 - 3(\frac{y}{c}\Big)^4 + \Big(\frac{y}{c}\Big)^6, \text{ ja } |y| \leq c \\ 1, \text{ ja } |y| > c, \end{cases}$$

tad paņemot $c = 0.70417$, $a = \sqrt{2}, b = 3/4$ iegūsim vienādojumu pret $\widehat{\sigma}_{MBT,n}(x)$ vērtības

$$\sum_{i=1}^{n-1} w_{n,i}(x)\chi_c\Big(\frac{Y_{i+1} - Y_i}{a\widehat{\sigma}_{MBT,n}(x)}\Big) - b = 0.$$

3. Praktiska daļa

3.1. Pielietojums simulētiem datiem

Iepriekšējā nodaļā tika aprakstīts, kā iegūt dispersijas novērtējumu caru dažādiem M - novērtējumiem. Šajā nodaļā aplūkosim šīs metodes pielietojumu simulētiem datiem. Sākumā, apskatīsim neparametriskās regresijas modeli

$$Y_i = g(x_i) + U_i\sigma(x_i), 1 \leq i \leq n, \tag{3.1}$$

kur $x_0 \leq x_1 \leq ... \leq x_n \leq 1$, σ nezināma dispersijas funkcija, g nezināma regresijas funkcija, kļūdas ir $U_i \sim F_0$ un $U_1, U_2, ..., U_n$ ir iid. Pie tam $g(x) = 2\sin(4\pi x)$, $\sigma(x) = e^x$, $x_i = i/(n+1)$, $1 \leq i \leq n$ un kļūdas U_i ir sadalītas pēc $G(y)$

$$G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y),$$

kur $\Phi(y) \sim N(0,1), H(y) \sim C(0,4^2)$, $n = 100$, $\varepsilon = 0, 0.1, 0.2, 0.3, 0.35, 0.4$.

Klasiskajam un robust novērtējumam, izmantosim Nadaraya - Watsona svarus

$$w_{n,i} = \frac{K\Big(\frac{x-x_i}{h_n}\Big)}{\sum_{j=1}^{n-1} K\Big(\frac{x-x_j}{h_n}\Big)},$$

kur K ir standarts Gausa kodols. Mēs izvēlamies $h_n = 0.2$ priekš mūsu simulētiem datiem.

Lai ģenerētu $G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim C(0,4^2)$ datus izmantosim sekojošus apgalvojumus, alternatīvu gadījuma lielumu ģenerēšanai:

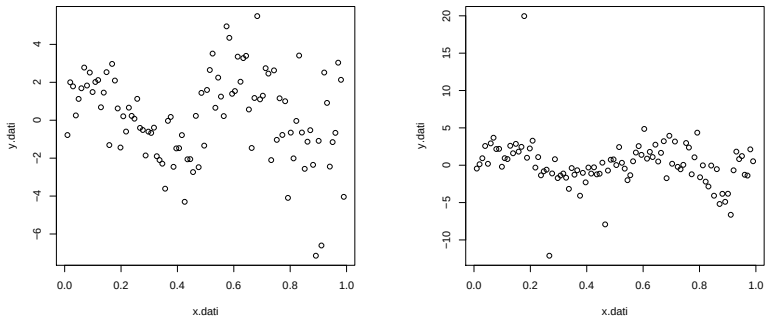
1. apgalvojums: Ja $X_1, X_2, ..., X_n$ ir vienādi neatkarīgi sadalīti (i.i.d.) gadījuma lielumi un $X_i \sim F$, tad $Y_1 = F(X_1), Y_2 = F(X_2), ..., Y_n = F(X_n)$ ir i.i.d. un pie tam $Y_i \sim v.s.[0,1]$

Pierādījums. $F_Y(y) = y$?

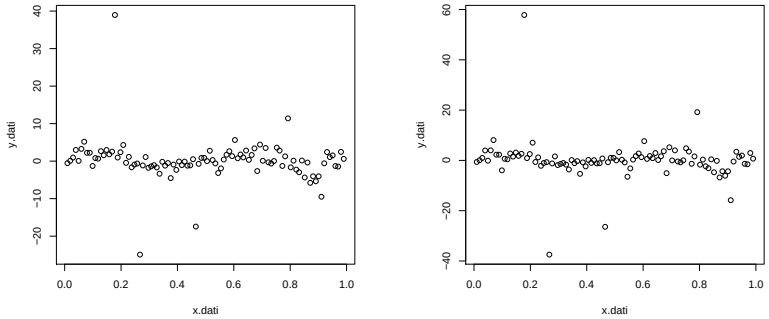
$$F_Y(y) = P(X \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y\square$$

2. apgalvojums: Ja doti $Y_1, Y_2, ..., Y_n; Y_i \sim v.s.[0,1]$, tad $X_1 = F^{-1}(Y_1), ..., X_n = F^{-1}(Y_n)$ ir i.i.d. un X_i ir sadalīts pēc F.

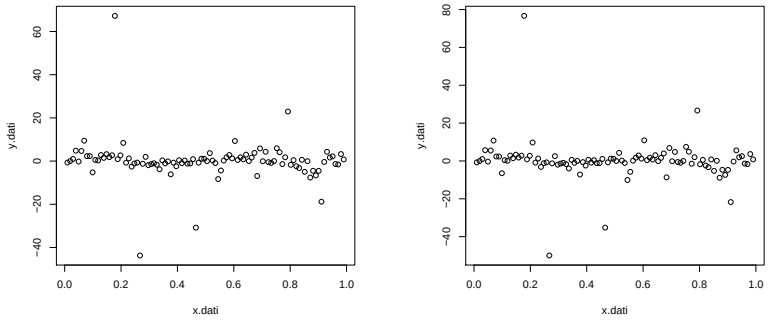
Paņemot $\varepsilon = 0, 0.1, 0.2, 0.3, 0.35, 0.4$, iegūstam datus



1. att. Datu grafiks pie $\varepsilon = 0$ un $\varepsilon = 0.1$.



2. att. Datu grafiks pie $\varepsilon = 0.2$ un $\varepsilon = 0.3$.



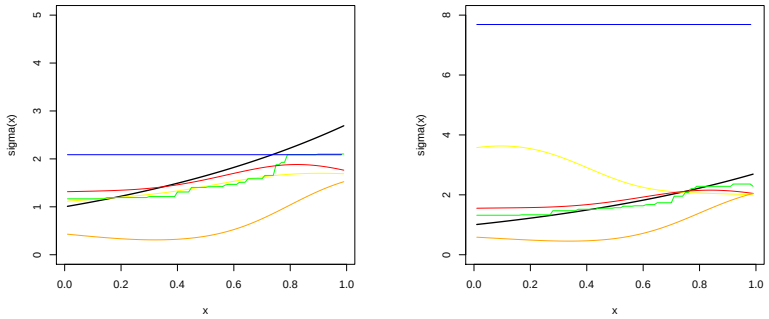
3. att. Datu grafiks pie $\varepsilon = 0.35$ un $\varepsilon = 0.4$.

Redzams, ka simulētiem datiem dispersija ir konstante ar izlēcējiem.

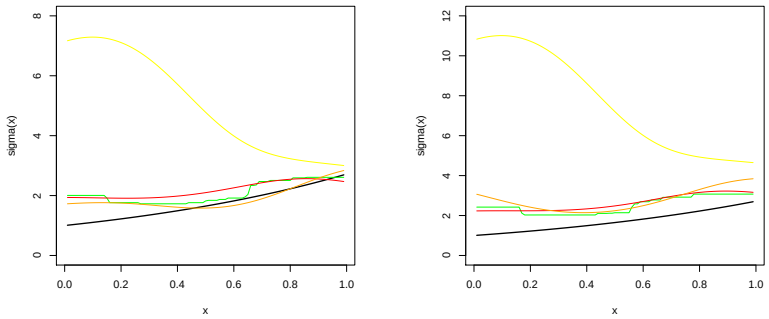
No 1.*att* līdz 3.*att* ir kļūst skaidrs, ka palielinot ε vērtību datiem simulējis vairāk izlēcēji. Šī procesa uzvedību uzdot loceklis $1 - \varepsilon$.

Uzzīmēsim īstas dispersijas funkcijas grafiku e^x simulētiem datiem $G(y) = (1 - \varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim C(0, 4^2)$ ar $\widehat{\sigma}_{RICE,n}(x)$, $\widehat{\sigma}_{MSD,n}(x)$, $\widehat{\sigma}_{MBT,n}(x)$,

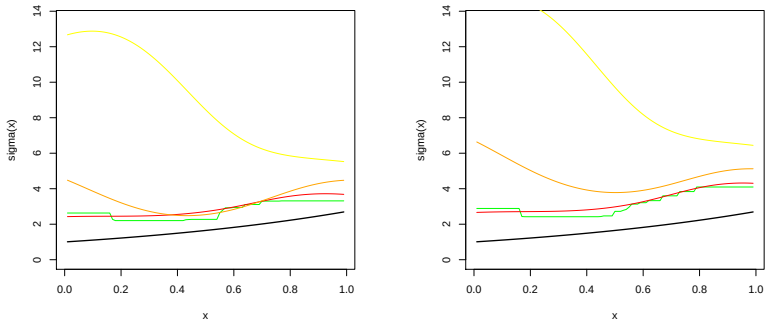
$\widehat{\sigma}_n(x)$ un $\widehat{\sigma}_{const,n}(x)$.



4. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0$ un $\varepsilon = 0.1$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, $\widehat{\sigma}_{const,n}(x)$ zila un $\widehat{\sigma}_n(x)$ oranža.



5. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0.2$, $\widehat{\sigma}_{const,n}(x) = 27.63$ un $\varepsilon = 0.3$, $\widehat{\sigma}_{const,n}(x) = 63.18$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana un $\widehat{\sigma}_n(x)$ oranža.



6. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0.35$, $\widehat{\sigma}_{const,n}(x) = 86.93$ un $\varepsilon = 0.4$, $\widehat{\sigma}_{const,n}(x) = 114.72$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana un $\widehat{\sigma}_n(x)$ oranža.

Vislabāk aproksimē īsto dispersiju $\widehat{\sigma}_{MSD,n}(x)$, jo tā līnija iet pietiekoši tuvu un dažās vietās sakrīt ar $exp(x)$ līniju. Vissliktāk aproksimē $\widehat{\sigma}_{const,n}(x)$, jo zila līnija iet ļoti tālu no īstas dispersijas līnijas.

Publikācija ir dota kļūda vērtīb

$$\widehat{ISEL} = \frac{1}{n}\sum_{i=1}^{n}\Big(log\Big(\frac{\widehat{\sigma}_n^j(x_i)}{\sigma(x_i)}\Big)\Big)^2,$$

kur $\widehat{\sigma}_n^{(j)}$ ir dispersijas novērtējums pēc j-tas metode($\widehat{\sigma}_{RICE,n}(x)$, $\widehat{\sigma}_{MSD,n}(x)$, $\widehat{\sigma}_{MBT,n}(x)$, $\widehat{\sigma}_n(x)$ vai $\widehat{\sigma}_{const,n}(x)$). Salīdzināsim, publikācijā dotas kļūdas vērtības ar mūsu iegūtiem pēc simulācijas.

1. tabula: $\widehat{ISEL}$ vērtība, kad $G(y) = (1-\varepsilon)\Phi(y)+\varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim C(0,4^2)$

novērtējums	$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 0.20$	$\varepsilon = 0.3$	$\varepsilon = 0.35$	$\varepsilon = 0.4$
$\widehat{\sigma}_{RICE,n}(x)$	0.021	3.89	6.63	8.70	9.61	10.43
$\widehat{\sigma}_{MSD,n}(x)$	0.036	0.074	0.20	0.47	0.67	0.92
$\widehat{\sigma}_{MBT,n}(x)$	0.052	0.082	0.18	0.36	0.49	0.65

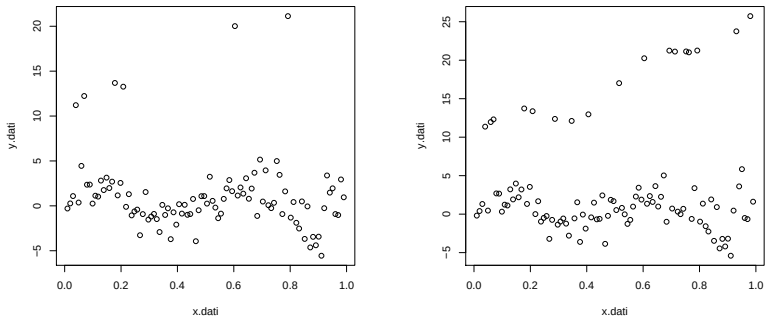
Veicot simulācijas ar programmu R tika iegūta sekojoša tabula:

2. tabula: $\widehat{ISEL}$ vērtība, kad $G(y) = (1-\varepsilon)\Phi(y)+\varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim C(0,4^2)$

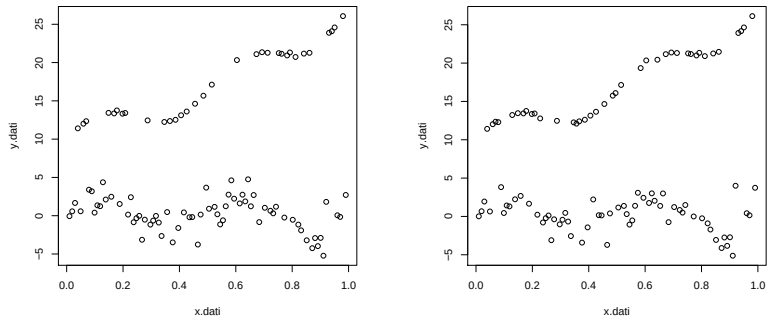
novērtējums	$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 0.20$	$\varepsilon = 0.3$	$\varepsilon = 0.35$	$\varepsilon = 0.4$
$\widehat{\sigma}_{RICE,n}(x)$	0.043	0.49	1.51	2.56	3.057	3.54
$\widehat{\sigma}_{MSD,n}(x)$	0.023	0.011	0.085	0.20	0.28	0.41
$\widehat{\sigma}_{MBT,n}(x)$	0.028	0.037	0.11	0.23	0.33	0.46
$\widehat{\sigma}_{n}(x)$	0.13	0.068	0.055	0.13	0.20	0.39
$\widehat{\sigma}_{const,n}(x)$	0.14	2.45	8.029	13.38	15.80	18.081

Abas tabulas rezultāti atšķiras, bet kopējā tendence saglabājas. Palielinot ε vērtību, palielinās arī $\widehat{ISEL}$ vērtība. Jo mazāka ir $\widehat{ISEL}$ vērtība, jo labāk. Redzams, ka vismazākās vērtības ir $\widehat{\sigma}_{MSD,n}(x)$ un $\widehat{\sigma}_{MBT,n}(x)$, bet vislielākās $\widehat{\sigma}_{RICE,n}(x)$ un $\widehat{\sigma}_{const,n}(x)$. To arī apstiprina iepriekš apskatītais 3.*att*.

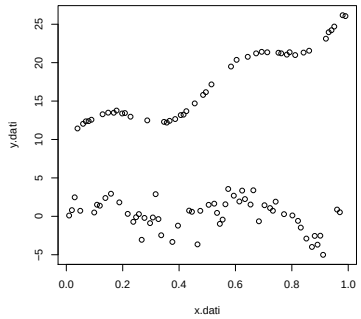
Simulēsim datus $G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim N(\mu, \sigma^2)$, kur $\sigma = 0.1$



7. att. Datu grafiks pie $\varepsilon = 0.1$ un $\varepsilon = 0.2$, $\mu = 10$.

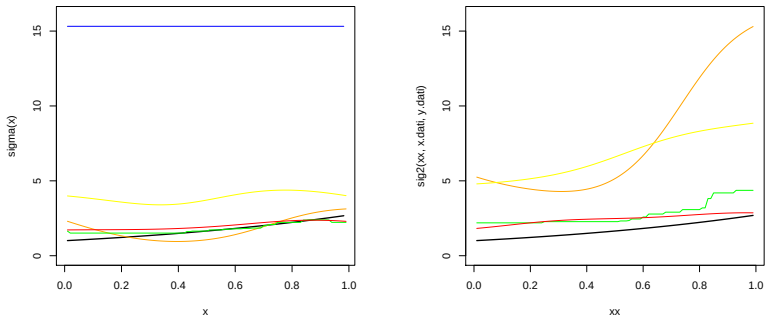


8. att. Datu grafiks pie $\varepsilon = 0.3$ un $\varepsilon = 0.35$, $\mu = 10$.

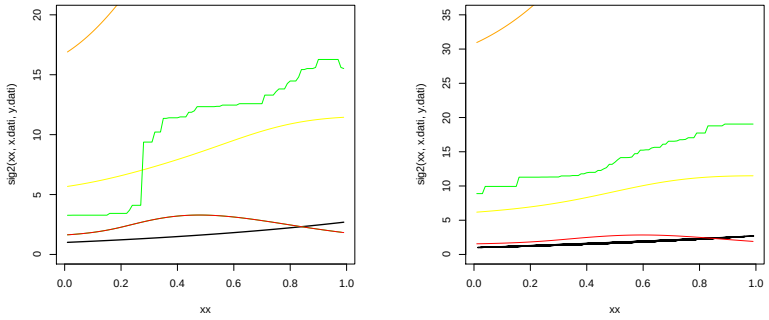


9. att. Datu grafiks pie $\varepsilon = 0.4$, $\mu = 10$.

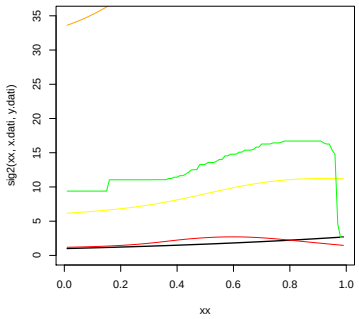
Uzzīmēsim īstas dispersijas funkcijas grafiku e^x simulētiem datiem $G(y) = (1 - \varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim N(\mu, \sigma^2)$ ar $\widehat{\sigma}_{RICE,n}(x)$, $\widehat{\sigma}_{MSD,n}(x)$, $\widehat{\sigma}_{MBT,n}(x)$, $\widehat{\sigma}_n(x)$ un $\widehat{\sigma}_{const,n}(x)$.



10. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0.1$ un $\varepsilon = 0.2$, $\mu = 10$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, $\widehat{\sigma}_n(x)$ oranža, $\widehat{\sigma}_{const,n}(x)$ zila, pie $\varepsilon = 0.2$ $\widehat{\sigma}_{const,n}(x) = 48.023$.



11. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0.3$ un $\varepsilon = 0.35$, $\mu = 10$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, $\widehat{\sigma}_n(x)$ oranža, pie $\varepsilon = 0.3$ $\widehat{\sigma}_{const,n}(x) = 80.11$, $\varepsilon = 0.35$ $\widehat{\sigma}_{const,n}(x) = 86.26$.



12. att.: Simulētiem datiem, dispersijas novērtējums, kā funkcijas pie $\varepsilon = 0.4$, $\mu = 10$. Melna līnija ir īsta dispersija $exp(x)$, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, $\widehat{\sigma}_{n}(x)$ oranža, $\widehat{\sigma}_{const,n}(x) = 80.11$.

Izlasei $G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim N(\mu, \sigma^2)$ publikācijā $\widehat{ISEL}$ vērtības ir šādas

3. tabula: $\widehat{ISEL}$ vērtība, kad $G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim N(\mu, \sigma^2)$

μ	novērtējums	$\varepsilon = 0.1$	$\varepsilon = 0.20$	$\varepsilon = 0.3$	$\varepsilon = 0.35$	$\varepsilon = 0.4$
10	$\widehat{\sigma}_{RICE,n}(x)$	1.35	2.060	2.45	2.57	2.66
	$\widehat{\sigma}_{MSD,n}(x)$	0.11	0.39	1.14	1.54	1.79
	$\widehat{\sigma}_{MBT,n}(x)$	0.099	0.20	0.32	0.37	0.39
100	$\widehat{\sigma}_{RICE,n}(x)$	11.53	13.74	14.90	15.13	15.34
	$\widehat{\sigma}_{MSD,n}(x)$	0.12	0.83	5.24	6.38	8.33
	$\widehat{\sigma}_{MBT,n}(x)$	0.11	0.30	1.00	1.23	1.64
1000	$\widehat{\sigma}_{RICE,n}(x)$	32.41	36.10	37.83	38.34	38.68
	$\widehat{\sigma}_{MSD,n}(x)$	0.15	1.20	10.92	16.82	21.56
	$\widehat{\sigma}_{MBT,n}(x)$	0.12	0.60	3.092	5.25	7.44

Veicot simulāciju ar programmu R tika iegūta sekojoša tabula:

4. tabula: $\widehat{ISEL}$ vērtība, kad $G(y) = (1-\varepsilon)\Phi(y) + \varepsilon H(y)$, $\Phi(y) \sim N(0,1)$, $H(y) \sim N(\mu, \sigma^2)$

μ	novērtējums	$\varepsilon = 0.1$	$\varepsilon = 0.20$	$\varepsilon = 0.3$	$\varepsilon = 0.35$	$\varepsilon = 0.4$
10	$\widehat{\sigma}_{RICE,n}(x)$	0.79	1.91	2.68	2.84	2.78
	$\widehat{\sigma}_{MSD,n}(x)$	0.026	0.25	3.03	4.46	4.18
	$\widehat{\sigma}_{MBT,n}(x)$	0.063	0.18	0.29	0.15	0.097
	$\widehat{\sigma}_{n}(x)$	0.11	2.023	9.61	12.24	12.71
	$\widehat{\sigma}_{const,n}(x)$	5.050	11.45	15.16	15.74	15.43
100	$\widehat{\sigma}_{RICE,n}(x)$	9.62	13.54	15.61	16.025	15.81
	$\widehat{\sigma}_{MSD,n}(x)$	0.026	0.25	13.88	19.95	19.04
	$\widehat{\sigma}_{MBT,n}(x)$	0.063	0.24	0.53	0.55	0.63
	$\widehat{\sigma}_{n}(x)$	14.22	41.35	62.14	66.84	68.073
	$\widehat{\sigma}_{const,n}(x)$	45.049	63.91	72.52	74.042	72.93
1000	$\widehat{\sigma}_{RICE,n}(x)$	29.20	35.82	39.14	39.79	39.43
	$\widehat{\sigma}_{MSD,n}(x)$	0.026	0.30	33.48	45.84	44.16
	$\widehat{\sigma}_{MBT,n}(x)$	0.063	0.26	0.78	1.47	2.10
	$\widehat{\sigma}_{n}(x)$	71.046	125.92	156.034	163.41	165.33
	$\widehat{\sigma}_{const,n}(x)$	128.11	158.85	172.22	174.60	172.80

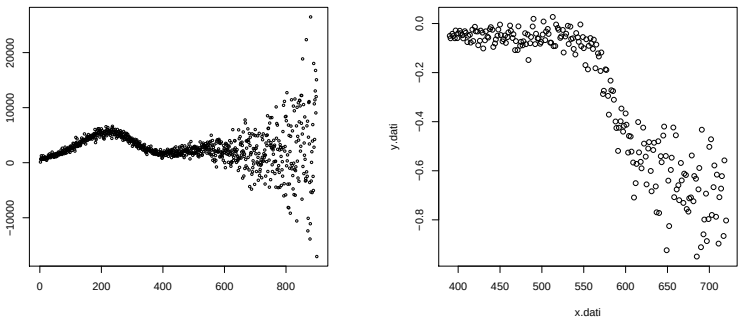
3.2. Pielietojums reāliem datiem

Pielietosim visas piecas aplūkotas metodes reāliem datiem CMB un LIDAR, kurus ņemam no Larry Wasserman grāmatas [1].

CMB ir kosmiskas radiācijas dati. Kuri rodas lielo sprādzienu rezultātā. X_i dati attēlo temperatūras fluktuācijas frekvenci un Y_i dati reprezentē fluktuācijas spēku katrā frekvencē. Fluktuācija ir fizikāla lieluma nejauša novirze no vidējās vērtības. Fluktuācija parādās procesos, kas pakļauti statistikas likumiem (molekulu, atomu, elektronu u. c. daļiņu kustība dažādās sistēmās). Daudzas fizikālas parādības var izskaidrot tikai ar fluktuācijas palīdzību, piemēram, Brauna kustību, gaismas molekulāro izkliedi, kas nosaka debesu zilo krāsu.

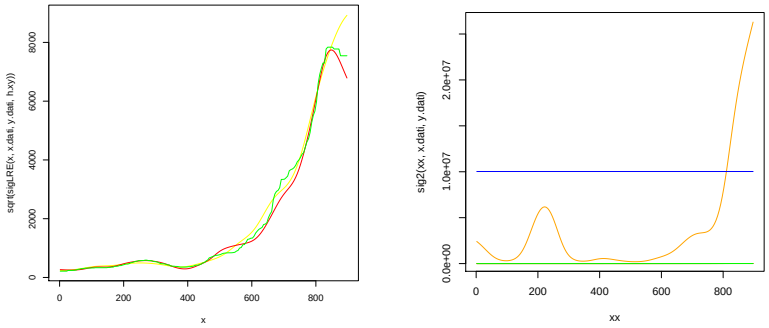
LIDAR dati ir iegūti no gaismas detektēšanas un apgabalu noteikšanas. LIDAR dati tiek lietoti, lai noteiktu piesārņojumu. X_i dati attēlo mērījuma distanci un Y_i dati reprezentē mērījuma ātrumu.

CMB un LIDAR visu datu izkliedes grafiks redzams 13. attēlā. No tā varam secināt, ka starp visas izlases atlikumiem pastāv neliela funkcionāla sakarība. CMB datiem dispersija pēc 400. datu punkta strauji pieaug. CMB izkliedes grafiks pirmajiem 400 datiem ir ar konstantu dispersiju, kas redzams 13. attēlā, bet par visiem atlikumiem to nevar pateikt.

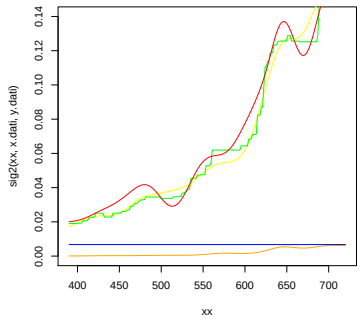


13. att. CMB un LIDAR dati.

Savukārt dispersijas funkcijas CMB un LIDAR datiem ir sekojošas:



14. att.: CMB datiem, dispersijas novērtējums, kā funkcijas, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, pie, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_n(x)$ oranža un $\widehat{\sigma}_{const}(x)$ ir zila.



15. att.: LIDAR datiem, dispersijas novērtējums, kā funkcijas, $\widehat{\sigma}_{MSD,n}(x)$ zaļa, $\widehat{\sigma}_{MBT,n}(x)$ sarkana, pie, $\widehat{\sigma}_{RICE,n}(x)$ dzeltena, $\widehat{\sigma}_n(x)$ oranža un $\widehat{\sigma}_{const}(x)$ ir zila.

Secinājumi

Darbā tika konstruēti dispersijas novērtējumi simulētiem un reāliem datiem, izmantojot dažādas pieejas. Tika apskatīti dispersijas lokālie M - novērtējumi ($\widehat{\sigma}_{MSD,n}(x)$, $\widehat{\sigma}_{MBT,n}(x)$, $\widehat{\sigma}_{RICE,n}(x)$) salīdzinājumā ar Yu un Jones [2] piedāvāto metodi $\widehat{\sigma}_n(x)$. Simulētiem datiem pie dažāda sadalījuma H izvēles, dispersijas funkciju labāk aproksimē lokālie M - novērtējumi. Tas izskaidrojams, ar to, ka lokālajos M novērtējumos tiek izmantota robusta statistika, bet $\widehat{\sigma}_n(x)$ tā netiek izmantota. Apskatot dispersijas novērtējumu ar konstanti $\widehat{\sigma}_{const}(x)$, esam pārliecinājušies, ka tā neattaisno sevi, jo gan simulētiem datiem, gan reāliem, tā atrodas pārāk tālu no īstās dispersijas līnijas.

Dispersijas funkcijas $\sigma(x)$ novērtējums pēc robusta ir svarīga problēma neparametriskās regresijas analīzē. Viennozīmīgi no salīdzinātajām M - novērtējuma metodēm $\widehat{\sigma}_{MSD,n}(x)$ ir vislabākā, jo visos simulētajos datu grafikos tā atrodas tuvāk īstajai dispersijas funkcijai nekā parējās līnijas un kļūdas $\widehat{ISEL}$ vērtības ir vismazākās.

Runājot par lokāliem M - novērtējumiem, jāpievērš uzmanību gludinošā parametra h un svaru $w_i(x)$ izvēlei. Simulētiem datiem mēs ņemam $h = 0.2$, tā kā bija apskatīts publikācijā. Reāliem datiem h izvēlas ar plug-in metodes palīdzību. Potenciāli h var izvēlēties arī pēc citām metodēm un paskatīties vai rezultāti uzlabosies. Ņemot svarus $w_i(x)$ ar Nadaraya-Watsona metodi redzam, ka galos funkcijas uzvedas dīvaini. Šo trūkumu var izlabot ņemot polinomiālus svarus.

Darba turpinājumā varētu nodarboties ar ticamības joslas konstruēšanu un lokālo M - novērtējumu uzlabojumu, ņemot citu χ funkciju, gludinošo parametru h un svarus $w_i(x)$ ar citiem paņēmieniem.

Izmantotā literatūra un avoti

[1] L. Wasserman. All of nonparametric statistics. 99:80–120, 2006.

[2] M. Jones. K. Yu. Likelihood-based local linear estimation of the conditional variance function. *Journal of the American Statistical Association*, 99:139–144, 2004.

[3] M. Levine. L. Brown. Varinace estimation in nonparametric regression via the difference sequence method. *Annals of Statistics*, 35:2219–2232, 2007.

[4] J. Rice. Bandwidth choice for nonparametric regression. *Annals of statistics*, 12:1215–1230, 1984.

[5] J. Meloche G. Boente, R. Fraiman. Robust plug-in bandwidth estimators in nonparametric regression. *Journal of Statistical Planning and Inference*, 57:109–142, 1997.

[6] T. Lee J. Hanning. Robust sizer for exploration of regression structures and outliers detection. *Journal of Computational and Graphical Statistics*, 15:101–117, 2006.

[7] T. Gasser W. Hardle. Robust nonparametric function fitting. *Journal of the Royal Statistical Society, Series B 46*, 9:42–51, 1984.

[8] A. Tsybakov W. Hardle. Robust nonparametric regression with simultaneous scale curve estimation. *Annals of Statistics*, 25:443–456, 1988.

[9] R. Fraiman. G. Boente. Robust nonparametric regression estimation for dependent observations. *Annals of statistics*, 17:1242–1256, 1989.

[10] E. Ronchetti. E. Cantoni. Resistant selection of the smoothing parameter for smoothing splines. *Statistics and Computing*, 11:141–146, 2001.

[11] D. Leung. Cross-validation in nonparametric regression with outliers. *Annals of Statistics*, 33:2291–2310, 2005.

[12] P. Huber. Robust estimation of a location parameter. *Annals of Statistics*, 35:73–101, 1964.

[13] D. Titterington P. Hall, J. Kay. Asymptotically optimal difference-based estimation of variance in nonparamteric regression. *Biometrika*, 77:521–528, 1990.

[14] R. Zamar I. Ghement, M. Ruiz. Robust estimation of error scale in nonparametric regression models. *Journal of Statistical Planning and Inference*, 138:3200–3216, 2008.

Programmas R kods

```
n<-100 #izlases apjoms

x.dati<-c(100) #definejam vektoru ar 100 koordinatem

x.dati<-1:n/(n+1) #definejam kopas X elementus

g<-function(x) #regresijas f-jas g(x) defineshana
{
2*sin(4*pi*x)
}

g(x.dati)

sigma<-function(x) #dispersijas f-jas sigma(x) defineshana
{
exp(x)
}

sigma(x.dati)

#atlukumu Ui genereshana
set.seed(1)

#H - koshi sadalijums ar mju=0 un sigma^2 = 4

#Yi vertibas defineshana
eps<-0 #parametrs epsilon
eps<-0.1
eps<-0.2
eps<-0.3
eps<-0.35
eps<-0.4

n<-100
x<-runif(n,0,1)
Gc<-c()
for (i in 1:n)
{
Gc[i]<-uniroot(function(y) (1-eps)*pnorm(y,0,1)+eps*pcauchy(y,0,4)-x[i],c(-500,500))$root
}

y.dati<-g(x.dati)+Gc*sigma(x.dati) #lidz ar to ir ieguti x un y dati

plot(x.dati,y.dati)


##############################################################################
```

```
#1. metode
#Novertejeam sigma, ar Rice metodi, ka f-ju no x: (sigma^2=(sigma(x))^2)
#Rice metodee datiem jabut sakartotiem
a<-sqrt(2)
b<-1

library(KernSmooth)
#Svaru novertejums izmantojot Nadarya-Wats, zinams, ka publikacija glud. parametrs h ir 0.2
#dpill ir prognoze ar neparametrisku f-ju, kura dod optimalo h-vertibu
dpill(x.dati,y.dati) #iebuveta f-ja noverteja joslas platumu h = 0.0289, pie eps=0
h.xy<-0.2 #ar h=0.2 labak aproksime nevis ar h garumu iegutu no dpill iebuvetas f-jas

#Nadarya-Watsona kodolu novertejums
NW.svari<-function(x,x.dat,y.dat,h,i) #nem katro i-to
{
dnorm((x-x.dat[i])/h)/sum(dnorm((x-x.dat)/h))
}
NW.svari<-Vectorize(NW.svari,vectorize.args="x") #novertejam f-ju r(x)
x<-seq(min(x.dati),max(x.dati),by=0.01)

#novertesim pashu f-ju sigma(x), ta ka sigma(x) defineta publikacija ar Rice palidzibu

NW.svari2<-Vectorize(NW.svari,vectorize.args="i")
sigLRE<-function(x,x.dat,y.dat,h)
{
n<-length(x.dat)
sum(NW.svari2(x,x.dat,y.dat,h,1:(n-1))*((diff(y.dat))/a)^2)
}
x<-seq(min(x.dati),max(x.dati),by=0.01)
sigLRE<-Vectorize(sigLRE,vectorize.args="x")

plot(x,sigma(x),ylim=c(0,8)) #uzzimejam istas disperijas(exp(x)) grafiku ar musu ieguto
lines(x,sqrt(sigLRE(x,x.dati,y.dati,h.xy)),col="yellow")#novertetas sigma vertibas izmantojot LRE
#katram xi bus sava sigma

##############################################################################
#2.metode

a<-sqrt(2)
b<-1/2

Q<-qnorm(0.75,0,1) #75 % kvantile

I<-c(99) #vieniba vektors
for (i in 1:length(y.dati)-1)
I[i]<-1

O<-c(99) #nulles vektors
```

```
for (i in 1:length(y.dati)-1)
O[i]<-0

Hic<-function(yy,sig,i)
{
yy<-diff(y.dati)/a
{if (abs(yy[i]/sig)>Q)
I[i]
else
O[i]}
}

Hic2<-Vectorize(Hic,vectorize.args="i") #katrai i-tai vertibai bus sava f-jas Hic vertiba

h.xy<-0.2

#Nadarya-Watsona kodolu novertejums
NW.svari<-function(x,x.dat,y.dat,h,i) #nem katro i-to
{
dnorm((x-x.dat[i])/h)/sum(dnorm((x-x.dat)/h))
}
NW.svari<-Vectorize(NW.svari,vectorize.args="x")
x<-seq(min(x.dati),max(x.dati),by=0.01)

#novertesim pashu f-ju sigma(x), ta ka sigma(x) defineta publikacija ar MT palidzibu

NW.svari2<-Vectorize(NW.svari,vectorize.args="i")
sigMSD<-function(x,x.dat,y.dat,h,sig)
{
n<-length(x.dat)
sum((NW.svari2(x,x.dat,y.dat,h,1:(n-1))*(Hic2(y.dat,sig,1:(n-1)))))-b
}
x<-seq(min(x.dati),max(x.dati),by=0.01)
sig2<-Vectorize(sigMSD,vectorize.args="sig")

#atrodisim visas saknes
nov.sig<-function(x)
{
uniroot(function(sig) sig2(x,x.dati,y.dati,h.xy,sig),c(0,10))$root
}
sigMSD<-Vectorize(nov.sig)

#katram xi bus sava sigma

lines(x,sigMSD(x),col="green") #uzzimejam novertetas disperijas grafiku


##############################################################################
#3.metode
```

```
#novertesim pashu f-ju sigma(x), ta ka sigma(x) defineta publikacija ar BT f-ju
c<-0.70417
a<-sqrt(2)
b<-0.75

I<-c()
for (i in 1:length(y.dati)-1)
I[i]<-1

Hic<-function(y.dati,sig,i)
{
yy<-diff(y.dati)/a
{if (abs(yy[i]/sig)<=c)
3*(yy[i]/(sig*c))^2-3*(yy[i]/(sig*c))^4+(yy[i]/(sig*c))^6
else
I[i]}
}

Hic2<-Vectorize(Hic,vectorize.args="i") #katrai i-tai vertibai bus sava f-jas Hic vertiba

h.xy<-0.2

#Nadarya-Watsona kodolu novertejums
NW.svari<-function(x,x.dat,y.dat,h,i) #nem katro i-to
{
dnorm((x-x.dat[i])/h)/sum(dnorm((x-x.dat)/h))
}
NW.svari<-Vectorize(NW.svari,vectorize.args="x")
x<-seq(min(x.dati),max(x.dati),by=0.01)

#novertesim pashu f-ju sigma(x), ta ka sigma(x) defineta publikacija ar BT palidzibu

NW.svari2<-Vectorize(NW.svari,vectorize.args="i")
sigMBT<-function(x,x.dat,y.dat,h,sig)
{
n<-length(x.dat)
sum((NW.svari2(x,x.dat,y.dat,h,1:(n-1))*(Hic2(y.dat,sig,1:(n-1)))))-b
}
x<-seq(min(x.dati),max(x.dati),by=0.01)
sigMBT<-Vectorize(sigMBT,vectorize.args="sig")

#atrodisim visas saknes
nov.sig<-function(x)
{
uniroot(function(sig) sig2(x,x.dati,y.dati,h.xy,sig),c(0,10))$root
}
sigMBT<-Vectorize(nov.sig)
```

```
#katram xi bus sava sigma

lines(x,sigMBT(x),col="red") #uzzimejam novertetas disperijas grafiku



################################################################################
#4. metode
#uzskata sigma par konstanti (rice metode)

# Nadaraya-Watsona kodolu regresija
NW.reg<-function(x,x.dati,y.dati,h)
{
sum(dnorm((x-x.dati)/h)*y.dati)/sum(dnorm((x-x.dati)/h))
}
NW.reg<-Vectorize(NW.reg,vectorize.arg="x")
x<-seq(min(x.dati),max(x.dati),by=0.009)
library(KernSmooth)
h<-dpill(x.dati,y.dati)
h

#novertesim sigma kaa konstante ar rice
n<-length(y.dati)
sig1<-sum(diff(y.dati)^2)/2/(n-1) #atrodam dispersijas vertibu kaa konstante

t<-length(x)
sigCONST<-c()
for (i in 1:length(x))
sigCONST[i]<-sig1

lines(x,sigCONST,col="blue")

################################################################################
#ISEL

#prieksh muusu simuletiem datiem, kad H(y)=C(0,4^2)

x<-seq(min(x.dati),max(x.dati),by=0.0099)

t1<-sqrt(sigLRE(x,x.dati,y.dati,h.xy)) #visas sigma vertibas no 1. metodes
t2<-sigMSD(x) #visas sigma vertibas no 2. metodes
t3<-sigMBT(x) #visas sigma vertibas no 3. metodes
t4<-sigCONST ##visas sigma vertibas no 4. metodes

ISEL<-function(signov,x.dati,i)
{
n<-length(x.dati)
(1/(n))*sum((log((signov[i])/(sigma(x.dati)[i])))^2)
}
```

```
#sigma(x.dati) ir taa istas disperijas f-jas vertiba, t.i. sigma(x.dati)=exp(x.dati)
ISEL(t1,x.dati,1:n)

##############################################################################

#dispersijas novertejums prieksh lidar datiem
#vispirms atrodisim atlikumus, lai uz tiem uzzimetu dipserisjas funkcijas

dati<-read.table(file="lider.txt",header=T)

x.dati<-dati$range
y.dati<-dati$logratio
plot(x.dati,y.dati)

#Nadarya-Wats kodolu regresija, raksta ar roku

NW.reg<-function(x,x.dati,y.dati,h)
{
sum(dnorm((x-x.dati)/h)*y.dati)/
sum(dnorm((x-x.dati)/h))
}
NW.reg<-Vectorize(NW.reg,vectorize.args="x") #novertejam f-ju r(x)
x<-seq(min(x.dati),max(x.dati),by=1.5)
library(KernSmooth)
h<-dpill(x.dati,y.dati)
h
lines(x,NW.reg(x,x.dati,y.dati,h)) #uzzimejam regresiju likne datiem

rez<-y.dati-NW.reg(x,x.dati,y.dati,h) #no atlikumiem var redzet ka disp nav konstanta
plot(x.dati,rez)

##############################################################################3

#sigma ka funkcija no x pec Nonparametric gramatas
#ta ir taada funkcija ar kuru mees salidzinam parejos metodes, jebkuriem datiem
library(KernSmooth)

sig2<-function(x,x.dati,y.dati)
{
h.YX<-dpill(x.dati,y.dati)
z.dati<-log((y.dati-NW.reg(x.dati,x.dati,y.dati,h.YX))^2)
h.ZX<-dpill(x.dati,z.dati)
exp(NW.reg(x,x.dati,z.dati,h.ZX))
}
xx<-seq(min(x.dati),max(x.dati),by=1.5)
sig2<-Vectorize(sig2,vectorize.args="x")

plot(xx,sig2(xx,x.dati,y.dati),ylim=c(0,0.1),type="l") #ista dispersijas f-ja lidar datiem
```

darbs “Neparametriskās regresijas kļūdas dispersijas novērtēšana” 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: Maksims Korotejevs

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.

Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents: Dr.math. Sandra Vucāne

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)