LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

VIENLAICĪGĀS TICAMĪBAS JOSLAS NEPARAMETRISKAJĀ REGRESIJĀ

MAĢISTRA DARBS

Autors: **Nataļja Saveļjeva**

Stud. apl. MaSt020005

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2009

Anotācija

Darbā tika apskatīta vienlaicīgo ticamību joslu konstruēšana neparametriskajai regresijai, izmantojot San un Loader un rezidiju butstrapa metodes. Šī darba mērķis ir pārādīt, ka heteroskedastiskā gadījumā (ja kļūdu dispersija ir atkarīga no regresora), mežonīgā rezidiju butstrapa metode strādā labāk nekā klasiskā metode, kas balstās uz parametriskiem nosacījumiem par kļūdu sadalījuma veidu.

Atslēgas vārdi: neparametriskā regresija, vienlaicīgās ticamības joslas, mežonīgais butstraps

Abstract

In this work was investigated a constructing of the simultaneous confidence bands in nonparametric regression, using Sun and Loader and residual bootstrap methods. The purpose of this work is to show, that in case of heteroskedasticity (then errors depend on exogenous variables) wild residual bootstrap method works better than classical method, which is based on the type of error distribution.

Keywords: nonparametric regression, simultaneous confidence bands, wild bootstrap

Saturs

Ievads

Regresijas analīzē svarīgu vietu aizņem ticamības joslas, lai gan ļoti svarīgi konstruēt arī pašu regresijas funkcijas novērtējumu. Eksistē divu veidu ticamības joslas: punktveida un vienlaicīgās. Literatūrā pārsvarā tiek apskatītas punktveida ticamības joslas, kas tiek saistītas ar vienlaicīgo joslu konstruēšanu sarežģītību, tomēr vienlaicīgās ticamības joslas dod informāciju kādā intervālā pie noteiktā nozīmības līmeņa atrodas visa nezināmā regresijas funkcija.

San un Loader metode skaitās klasiskā metode neparametriskajā regresijā, kas balstās uz parametriskiem pieņēmumiem par kļūdu normalitāti. Šī metode tika pamatota samērā nesen, 1994. gadā [1]. Šī darba mērķis ir izanalizēt alternatīvu rezidiju butstrapu, kas sāka attīstīties no 1986. gada [2], kā arī salīdzināt to ar klasisko San un Loader metodi. Šī tēma ir aktuāla mūsdienās. Viens no pēdējiem darbiem ir Liu, Wei un Lin, Shan (2009) darbs par vienlaicīgajām ticamības joslām daudzdimensiju lineārajā regresijā, kā arī Zhibio Zhao un Wei Biao Wu (2008) darbs par vienlaicīgajām ticamības joslām neparametriskajā laikrindu regresijā.

Visbeidzot viens no darba mērķiem ir salīdzināt metodes, analizējot pārklājuma precizitātes ar simulāciju palīdzību, kā arī konstruēt ticamības joslas praktiskām datu problēmām. Vienlaicīgās ticamības joslas tika konstruētas uz CMB (cosmic microwave background radiation) datiem no Larry Wassermana grāmatas [3], kā arī uz veselības apdrošināšanas datiem, kas tika iegūti no apdrošināšanas kompānijas "ERGO Latvija Dzīvība".

Darbs sastāv no 6 nodaļām un pielikuma. 1.nodaļā definēti daži pamatjēdzieni, kas izmantoti darbā. 2.nodaļā ir aprakstīta neparametriskā regresija, kurai tiek konstruētas vienlaicīgās ticamības joslas. 3.nodaļā ir aprakstīta butstrapa metode, tajā skaitā rezidiju butstraps, pāru butstraps un rezidiju mežonīgais butstraps. 4.nodaļā ir aprakstītas vienlaicīgo ticamības joslu konstruēšanas metodes. 5.nodaļā, izmantojot veiktās simulācijas, ir salīdzinātas mežonīgā rezidiju butstrapa vienlaicīgo ticamības joslu konstruēšanas metode ar San un Loader metodi. 6.nodaļā darbā apskatītās metodes ir pielietotas reālai datu problēmai. Pēdējā nodaļā ir aprakstīti galvenie darbā iegūtie rezultāti un secinājumi. Izmantotās literatūras saraksts ir ievietots pēc pēdējās nodaļas. Lai analizētu un pētītu minētās metodes vienlaicīgo ticamības joslu konstruēšanai neparametriskajā regresijā, programmā R tika uzrakstītas vairākas datorprogrammas, dažas no tām ir pievienotas pielikumā.

Apzīmējumu saraksts

$X_n = o(a_n)$ $\lim\limits_{n\to\infty} \frac{X_n}{a_n} = 0$

$X_n = O(a_n)$ $|X_n/a_n|$ ir ierobežots visiem lieliem n

$X_n \xrightarrow{p} X$ konverģence pēc varbūtības

$X_n \xrightarrow{d} X$ konverģence pēc sadalījuma

$X_n = o_p(a_n)$ $\frac{X_n}{a_n} \xrightarrow{p} 0$

$X_n = O_p(a_n)$ $\left|\frac{X_n}{a_n}\right|$ ir ierobežots varbūtībā visiem lieliem n

1. Pamatjēdzieni un definīcijas

Šajā nodaļā apskatīsim dažus darbā izmantotos pamatjēdzienus. Jēdzienu definēšanā izmantoti literatūras avoti [3], [4].

Pieņemsim, ka $X_1, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi un f ir blīvuma funkcija.

Definīcija 1. *Ja $\hat{f}_n(x)$ ir funkcijas $f(x)$ novērtējums punktā x, tad kvadrātiskās kļūdas zaudējuma funkcija ir*

$$L(f(x), \hat{f}_n(x)) = (f(x) - \hat{f}_n(x))^2.$$

Definīcija 2. *Zaudējuma funkcijas vidējo vērtību sauc par risku*

$$R(f(x), \hat{f}_n(x)) = E(L(f(x), \hat{f}_n(x))) = (E(\hat{f}_n(x)) - f(x))^2 + D(\hat{f}_n(x)).$$

Definīcija 3. *Integrētais risks tiek definēts sekojoši*

$$R(f(x), \hat{f}_n(x)) = \int R(f(x), \hat{f}_n(x))dx.$$

Definīcija 4. *Krosvalidāciju riska novērtējums ir*

$$\hat{J}(h) = \int \hat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i),$$

kur $\hat{f}_{-i}(X_i)$ ir blīvuma funkcijas novērtējums, ko iegūst pēc i - tā novērojuma atmešanas. $\hat{J}(h)$ sauc par novērtēto risku jeb krosvalidācijas skores funkciju.

2. Neparametriskā regresija

Šajā nodaļā tiek dots neparametriskās regresijas teorijas apskats, kas vēlāk tiks izmantots vienlaicīgo ticamības joslu konstruēšanai. Neparametriskās metodes nodarbojas ar standarta statistiskajām problēmām gadījumā, ja nosacījumi par normālo populācijas sadalījumu ir aizvietoti ar vispārējiem nosacījumiem par populāciju vai populācijas sadalījumu.

Pieņemsim, ka doti n neatkarīgi un vienādi sadalīti novērojumu pāri $(X_1, Y_1), \ldots, (X_n, Y_n)$. Starp rezultējošo mainīgo (atkarīgo) Y un neatkarīgo mainīgo X ir sekojoša savstarpēja sakarība

$$Y_i = r(X_i) + \epsilon_i, \;\; E(\epsilon_i) = 0, \;\; i = 1, ..., n, \tag{2.0.1}$$

kur r ir regresijas funkcija. Mērķis neparametriskajā regresijā ir novērtēt pašu funkciju r. $r(x)$ novērtējums tiek apzīmēts ar $\hat{r}_n(x)$. Tika pieņemts, ka dispersija $D(\epsilon_i) = \sigma^2$ nav atkarīga no X. Regresijas līkne tiek interpretēta kā Y vidējā vērtība pie dotas vērtības x , tas ir $r(x) = E(Y|X = x)$. Šajā nodaļā tiek apskatītas tādas lokālas regresijas metodes kā kodolu regresija un lokālā polinomu regresija. Šeit visi novērtējumi ir lineārie gludinātāji. Pirms iedziļināsimies neparametriskajā regresijā, vispirms īsi tiks apskatīta parametriskā pieeja.

2.1. Lineāras regresijas modelis

Tiek doti neatkarīgi un vienādi sadalīti dati $(X_1, Y_1), \ldots, (X_n, Y_n)$, kur $Y_i \in \mathbb{R}$ un $X_i = (X_{i1}, ..., X_{ip})^T \in \mathbb{R}^p$. Lineāras regresijas modelis tiek definēts kā

$$Y_i = r(X_i) + \epsilon_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \;\; i = 1, \ldots, n, \tag{2.1.1}$$

kur $E(\epsilon_i) = 0$, $D(\epsilon_i) = \sigma^2$ un $X_{i1} = 1$. Dizaina matrica $\mathbf{X}$ ir $n \times p$ matrica

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix}.$$

Vektoru kopa Υ sastāv no matricas $\mathbf{X}$ lineārām kombinācijām.

Pierakstīsim lineārās regresijas modeļa komponentes vektoru formā: $Y = (Y_1, ..., Y_n)^T$, $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ un $\beta = (\beta_1, ..., \beta_p)^T$. Tātad modeli (2.1.1) var pārrakstīt kā

$$Y = \mathbf{X}\beta + \epsilon.$$

Mazāko kvadrātu novērtējums $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$ ir vektors, kurš minimizē kvadrātu atlikumu summu

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

[5] Ja matricai $\mathbf{X}^T\mathbf{X}$ eksistē inversā matrica, tad labākais novērtējums ir

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

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

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

No tā seko, ka novērtēto vērtību $\mathbf{r} = (\hat{r}_n(X_1), ..., \hat{r}_n(X_n))^T$ var pierakstīt kā

$$\mathbf{r} = \mathbf{X}\hat{\beta} = LY,$$

kur

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

Matricu L sauc par cepures (hat) matricu, bet vektora $\hat{\epsilon} = Y - \mathbf{r}$ elementus par rezidijiem. Matrica L ir simetriska un idempotenta, tas ir, $L = L^T$ un $L^2 = L$. No tā seko, ka $\mathbf{r}$ ir Y projekcija uz telpu Υ , un parametru skaita p un matricas L attiecība ir tāda, ka

$$p = tr(L).$$

Dotam punktam $x = (x_1, ..., x_p)^T$

$$\hat{r}_n(x) = l(x)^T Y = \sum_{i=1}^{n} l_i(x) Y_i, \tag{2.1.2}$$

kur

$$l(x)^T = x^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T.$$

2.2. Lineārie gludinātāji

Definīcija 5. *Funkcijas r novērtējums $\hat{r}_n$ ir lineārs gludinātājs, ja katram x eksistē vektors $l(x) = (l_1(x), ..., l_n(x))^T$ tāds, ka*

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

Lineārās regresijas novērtējumi arī ir lineāri gludinātāji (skatīt (2.1.2)).

Definēsim novērtēto vērtību vektoru kā

$$\mathbf{r} = (\hat{r}_n(X_1), ..., \hat{r}_n(X_n))^T,$$

un $Y = (Y_1, ..., Y_n)^T$. Tad

$$\mathbf{r} = LY,$$

kur L ir $n \times n$ matrica, kur $i-$tajā rindiņa ir $l(X_i)^T$, tādējādi $L_{ij} = l_j(X_i)$. Un $i-$tās rindiņas elementi parada svarus, kas piešķirti Y_i, veidojot $\hat{r}_n(X_i)$.

Definīcija 6. *Matricu L sauc par gludināšanas vai cepures matricu. Matricas L $i-$to rindiņu sauc par efektīvo kodolu $r(X_i)$ novērtēšanai. Efektīvas brīvības pakāpes ir*

$$\nu = tr(L)$$

Ievērosim, ka $\sum_{i=1}^{n} l_i(x) = 1$. Ja $Y_i = c$ katram i, tad $\hat{r}_n(x) = c$.

2.3. Lokālā regresija

Pieņemsim, ka $x_i \in \mathbb{R}$ un regresijas modelis ir formā (2.0.1). Šajā sadaļā tiek apskatīts $r(x)$ novērtējums, kas tika iegūts, ņemot vidējo svērto no Y_i mainīgajiem, piešķirot lielākus svarus tiem punktiem, kuri ir blakus x. Vispirms apskatīsim kodolu regresijas novērtējumu.

Definīcija 7. *Pieņemsim, ka $h > 0$ pozitīvs skaitlis, kuru sauc par joslas platumu. Nadaraja-Watsona kodolu novērtējums ir definēts kā*

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

kur K ir kodols un $l_i(x)$ ir svari, kas tiek definēti kā

$$l_i(X) = \frac{K\left(\frac{x-X_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right)}.$$

Definīcija 8. *Par kodolu sauksim jebkuru gludu funkciju K tādu, ka $K(x) \geq 0$, $\int K(x)dx = 1$, $\int xK(x)dx = 0$ un $\int x^2 K(x)dx > 0$.*

Visbiežāk lietoti kodolu piemēri: boxcar kodols

$$K(x) = \begin{cases} \frac{1}{2}, \, |x| \geq 1 \\ 0, \text{ pretējā gadījumā} \end{cases}.$$

normālais jeb Gausa kodols

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

Epaņečņikova kodols

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), \, |x| \geq 1 \\ 0, \text{ pretējā gadījumā} \end{cases}.$$

Definīcija 9. *Ja K ir kodols un joslas platums h ir pozītīvs skaitlis, tad*

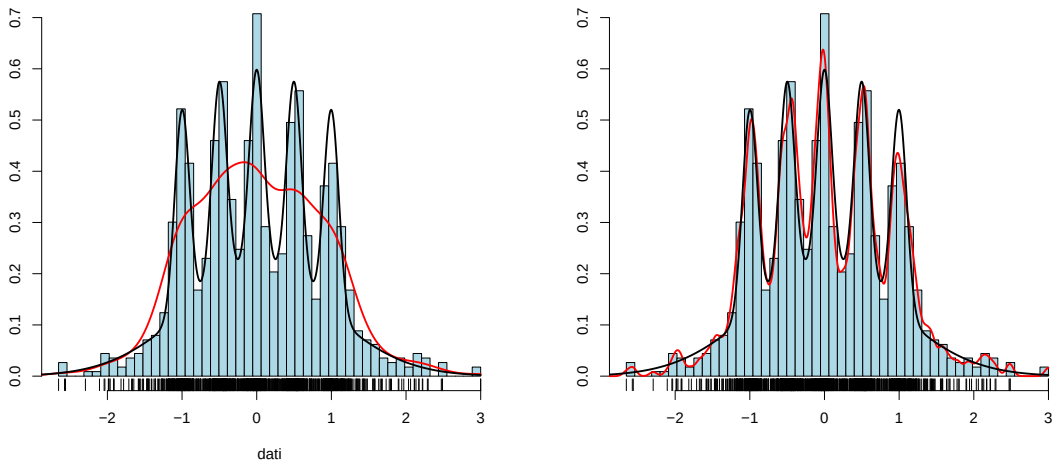
$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right)$$

ir kodolu blīvuma funkcijas novērtējums.

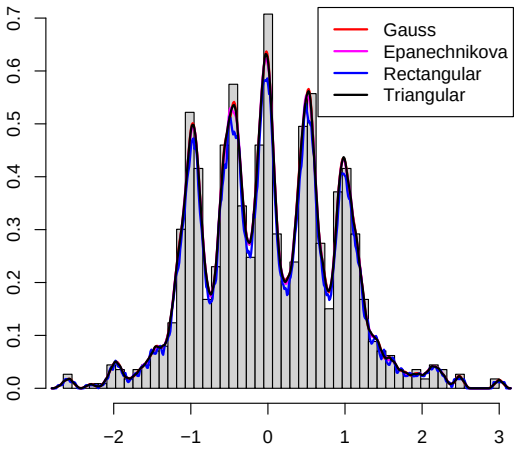
Kodola K izvēle nav tik svarīga, tā kā skaitliski ar dažādiem kodoliem iegūtie novērtējumi ir ļoti līdzīgi. Savukārt joslas platuma h izvēle ir būtisks jautājums, jo šis parametrs kontrolē gludināšanas pakāpi. Mazi joslas platumi dod ļoti negludus novērtējums, bet lielāki h dod gludākus novērtējumus. Ilustratīvi piemēri tiek atspoguļoti Attēlos 2.1. un 2.2.

Gludinātāji, kuri tiek izmantoti šajā darbā, ir atkarīgi no gludināšanas parametra h, tāpēc ir nepieciešama kaut kāda procedūra joslas platuma izvēlei. Definēsim risku jeb vidējo kvadrātisko kļūdu

$$R(h) = E\left(\frac{1}{n} \sum_{i=1}^{n} (\hat{r}_n(X_i) - r(X_i))^2\right).$$



2.1. att.: Datu blīvuma funkcijas kodolu novērtējums ar dažādiem joslas platumiem h pie vienāda kodola



2.2. att.: Datu blīvuma funkcijas novērtējums ar dažādiem kodoliem pie vienāda joslas platuma h

Ideāli būtu izvēlēties h, minimizējot $R(h)$, bet $R(h)$ ir atkarīgs no nezināmās funkcijas $r(x)$. Tāpēc minimizēsim $R(h)$ novērtējumu $\hat{R}(h)$. Pirmā ideja ir minimizēt vidējo kvadrātu atlikumu summu

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_n(X_i))^2,$$

lai novērtētu $R(h)$. Bet tas nav labs riska novērtējums, jo ir novirzīts un nepietiekoši nogludināts. Iemesls tam ir tāds, ka dati tiek izmantoti divreiz: lai novērtētu regresijas

funkciju un risku. Tāpēc novērtēsim risku, pielietojot "vienu-atstāt-ārā" krosvalidācijas funkciju.

Definīcija 10. *"Vienu-atstāt-ārā" krosvalidācijas funkcija (leave-one-out score function) tiek definēta kā*

$$CV = \hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{r}_{-i}(X_i))^2,$$

kur $\hat{r}_{-i}$ *ir novērtējums, kas iegūstams, izlaižot pāri* (X_i, Y_i).

Definīcija 11.

$$\hat{r}_{-i}(x) = \sum_{i=1}^{n} Y_j l_{j,(-i)}(x),$$

kur

$$l_{j,(-i)}(x) = \begin{cases} 0, & ja \quad j = i \\ \frac{l_j(x)}{\sum_{k\neq i} l_k(x)}, & ja \quad j \neq i. \end{cases}$$

Ievērosim, ka

$$E(Y_i - \hat{r}_{-i}(X_i))^2 = E(Y_i - r(X_i) + r(X_i) - \hat{r}_{-i}(X_i))^2 =$$

$$= \sigma^2 + E(r(X_i) - \hat{r}_{-i}(X_i))^2 \approx \sigma^2 + E(r(X_i) - \hat{r}_n(X_i))^2,$$

tādējādi

$$E(\hat{R}) \approx R + \sigma^2.$$

Tādā veidā krosvalidācijas metode dod gandrīz nenovirzītu novērtējumu. Tas ir acīmredzams, ka $\hat{R}(h)$ aprēķināšanas process būs laikietilpīgs, jo nāksies pārrēķināt novērtējumu pēc katra novērojuma atmešanas. Lineāram gludinātājam eksistē vienkāršotā formula $\hat{R}$ aprēķināšanai (skatīt [3]).

Apgalvojums 1. *Ja* $\hat{r}_n$ *ir lineārs gludinātājs, tad*

$$\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{r}_n(X_i)}{1 - L_{ii}}\right)^2, \tag{2.3.1}$$

kur $L_{ii} = l_i(X_i)$ *ir* $i-$*tais diagonālais gludināšanas matricas* L *elements.*

Tātad gludināšanas parametru h var izvēlēties, minimizējot funkciju $\hat{R}(h)$. Alternatīva metode krosvalidācijai joslas platuma izvēlei ir plug-in novērtējums [3], [6].

Nākamā teorēma demonstrē, kā joslas platums ietekmē novērtējumu.

Teorēma 1. *Risks Nadaraya-Watsona kodola novērtējumam ir*

$$R(\hat{r}_n, r) = \frac{h_n^4}{4}\left(\int x^2 K(x)dx\right)^2 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx + \tag{2.3.2}$$
$$+\frac{\sigma^2 \int K^2(x)dx}{nh_n}\int \frac{1}{f(x)}dx + o(nh_n^{-1}) + o(h_n^4).$$

Turklāt, ja $h_n \to 0$, $nh_n \to \infty$, $EY^2 < \infty$, $f(x) > 0$, *tad* $\hat{r}_n(X) \xrightarrow{p} r(X)$.

Pirmais saskaitāmais riska novērtējuma izteiksmē ir novirze kvadrātā, un otrais saskaitāmais ir dispersija. Lielumu

$$2r'(x)\frac{f'(x)}{f(x)},$$

kuru satur novirze, sauc par dizaina novirzi, jo tas ir atkarīgs no X sadalījuma. Tas nozīmē, ka novirze ir ļoti jūtīga pret X izvietojumu. Kodolu novērtējumiem ir lielas novirzes no īstā novērtējuma netālu no robežām, to sauc par robežu novirzi.

$R(\hat{r}_n, r)$ visātrāk konverģē, kad abi divi saskaitāmie, novirze kvadrātā un dispersija konverģē vienādi, tas ir $O(n^5) = O(\frac{1}{nh})$. Tā ir novirzes - dispersijas kompromisa problemātika. Tātad iegūtais optimālais joslas platums ir

$$h_* = \left(\frac{1}{n}\right)^{1/5}\left(\frac{\sigma^2 \int K^2(x)dx \int dx/f(x)}{(\int x^2K^2(x)dx)^2 \int \left(r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\right)^2 dx}\right)^{1/5}.$$

Tādējādi $h_* = O(n^{-1/5})$. Ievietojot optimālo novērtējumu h_* novērtējumā (2.3.2), var redzēt, ka risks samazinās ar ātrumu $O(n^{-4/5})$. Parametriskajām metodēm riska konverģences ātrums ir $O(n^{-1})$ [3]. Neparametriskās metodes nedaudz sliktāk strādā nekā parametriskās metodes. Bet parametriskās metodes strādā labi tikai tad, ja blīvuma funkcijas novērtējums ir labs, kamēr neparametriskās metodes strādā vienmēr.

2.4. Lokālā polinomu regresija

Kodolu novērtējuma problēma ir dizaina un robežu novirze, bet to var atrisināt ar lokālo polinomu regresiju.

Apskatīsim novērtējuma $a \equiv \hat{r}_n(x)$ izvēli, minimizējot kvadrātu summu $\sum_{i=1}^{n}(Y_i - a)^2$. Optimizācijas problēmas atrisinājums ir konstanta $\hat{r}_n(x) = \overline{Y}$, kura acīmredzot nav labs $r(x)$ novērtējums. Tālāk definēsim funkcijas svarus $w_i(x) = K((X_i - x)/h)$ un minimizēsim svērto kvadrātu summu

$$\sum_{i=1}^{n} w_i(x)(Y_i - a)^2.$$

Rezultātā iegūstam sekojošo atrisinājumu

$$\hat{r}_n(x) \equiv \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}.$$

Tātad Nadaraya-Watsona kodolu regresijas novērtējums ir faktiski regresijas lokālā aproksimācija ar konstanti. Tālāk uzlabosim novērtējumu, izmantojot lokālo polinomu ar kārtu p konstantes a vietā. Pieņemsim, ka x ir fiksēts punkts, kurā tiks novērtēta funkcija $r(x)$. Definēsim polinomu vērtībām u punkta x apkārtnē:

$$P_x(u; a) = a_0 + a_1(u - x) + \frac{a_2}{2!}(u - x)^2 + \ldots + \frac{a_p}{p!}(u - x)^p.$$

Nākamais solis ir $a = (a_0, \ldots, a_p)^T$ aproksimēšana, izvēloties $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_p)$ tādu, kas minimizē lokāli svērto kvadrātu summu

$$\sum_{i=1}^n w_i(x)(Y_i - P_x(X_i; a))^2. \tag{2.4.1}$$

Novērtējums $\hat{a}$ ir atkarīgs no fiksēta punkta x. Tātad funkcijas r lokālais novērtējums ir

$$\hat{r}_n(u) = P_x(u; \hat{a}).$$

Speciālajā gadījumā, kad $u = x$, iegūstam

$$\hat{r}_n(x) = P_x(x; \hat{a}) = \hat{a}_0(x).$$

Piezīme 1. *Ja $p = 0$, tad iegūstam kodolu novērtējumu. Ja $p = 1$, tad to sauc par lokālo lineāro regresiju.*

Lai atrastu $\hat{a}$, pārdefinēsim problēmu vektoru formā. Pieņemsim, ka dizaina matrica ir definēta kā

$$\mathbf{X}_x = \begin{pmatrix} 1 & X_1 - x & \ldots & \frac{(X_1-x)^p}{p!} \\ 1 & X_2 - x & \ldots & \frac{(X_2-x)^p}{p!} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n - x & \ldots & \frac{(X_n-x)^p}{p!} \end{pmatrix}.$$

Minimizējot izteiksmi (2.4.1), iegūstam svērto mazāko kvadrātu novērtējumu

$$\hat{a}(x) = (\mathbf{X}_x^T W_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T W_x Y.$$

Speciālajā gadījumā $\hat{r}_n(x) = \hat{a}_0(x)$ ir matricas $(\mathbf{X}_x^T W_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T W_x$ pirmā rindiņa sareizinināta ar Y. Tādējādi var nodefinēt sekojošu teorēmu (skatīt [3]).

Teorēma 2. *Lokālās polinomu regresijas novērtējums ir*

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

kur $l(x)^T = (l_1(x), ..., l_n(x))$,

$$l(x)^T = e_1^T (\mathbf{X}_x^T W_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T W_x,$$

$e_1 = (1, 0, ..., 0)^T$. *Šim novērtējumam matemātiskā cerība ir*

$$E(\hat{r}_n(x)) = \sum_{i=1}^{n} l_i(x) r(X_i)$$

un dispersija ir

$$D(\hat{r}_n(x)) = \sigma^2 \sum_{i=1}^{n} l_i(x)^2 = \sigma^2 \|l(x)\|^2.$$

Tātad šis novērtējums ir lineārs gludinātājs un joslas platumu var izvēlēties, minimizējot krosvalidācijas formulu (2.3.1).

Teorēma 3 (Lokālā lineārā gludināšana)**.** *[3] Ja* $p = 1$, *tad* $\hat{r}_n(x) = \sum_{i=1}^{n} l_i(x) Y_i$, *kur*

$$l_i(x) = \frac{b_i(x)}{\sum_{j=1}^{n} b_j(x)},$$

$$b_i(x) = K\left(\frac{X_i - x}{h}\right)(S_{n,2}(x) - (X_i - x) S_{n,1}(x))$$

un

$$S_{n,j}(X) = \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(X_i - x)^j, \ j = 1, 2.$$

Nākamā teorēma liecina par to, ka lokālā lineārā regresija ir labāka par kodolu regresiju.

Teorēma 4. *[3] Pieņemsim, ka* $Y_i = r(X_i) + \sigma(X_i)\epsilon_i$ *katram* $i = 1, \ldots, n$, $a \le X_i \le b$, *un* $X_1, \ldots, X_n$ *ir izlase no kāda sadalījuma ar tādu blīvuma funkciju* f, *ka* (i) $f(x) > 0$, (ii) f, r'' *un* σ^2 *ir nepārtraukti punkta* x *apkārtnē, un* (iii) $h_n \to 0$ *un* $nh_n \to \infty$. *Pieņemsim, ka* $x \in (a, b)$. *Dotiem* $X_1, \ldots, X_n$ *iegūstam, ka lokālajam lineārajam novērtējumam un kodolu novērtējumam dispersija ir*

$$\frac{\sigma^2(x)}{f(x) n h_n} \int K^2(u) du + o_P\left(\frac{1}{n h_n}\right).$$

Nadaraya-Watsona kodolu novērtējumam novirze ir

$$h_n^2 \left(\frac{1}{2} r''(x) + \frac{r'(x) f'(x)}{f(x)}\right) \int u^2 K(u) du + o_P(h^2),$$

kamēr lokālajam lineārajam novērtējumam ir asimptotiskā novirze

$$h_n^2 \frac{1}{2} r''(x) \int u^2 K(u) du + o_P(h^2).$$

Tādējādi lokālais lineārais novērtējums ir brīvs no dizaina novirzes. Robežu punktos a un b Nadaraya-Watson kodolu novērtējumam ir asimptotiskā novirze ar kārtu h_n, kamēr lokālajam lineārajam novērtējumam ir novirze ar kārtu h_n^2. Šajā nozīmē lokālā lineārā novērtēšana novērš robežu novirzi.

2.5. Dispersijas novērtēšana

Pieņemsim, ka novērojumu pāri $(X_1, Y_1), \ldots, (X_n, Y_n)$ ir neatkarīgi un vienādi sadalīti un apskatāmais modelis ir

$$Y_i = r(X_i) + \sigma(X_i)\epsilon_i, \quad i = 1, \ldots, n,$$

kur $E\epsilon_i = 0$ un $D\epsilon_i = 1$. Šajā sadaļā tiek apskatītas dažas metodes dispersijas $\sigma^2(X_i)$ novērtēšanai gadījumos, kad tā ir atkarīga no regresora un ir konstanta.

Nākamās divas teorēmas dod nenovirzītus novērtējumus parametram $\sigma^2(X_i) = \sigma^2$ (skatīt [3]).

Teorēma 5. *Pieņemsim, ka $\hat{r}_n(X)$ ir lineārs gludinātājs un*

$$\sigma^2 = \frac{\sum_{i=1}^n (Y_i - \hat{r}(X_i))^2}{n - 2\nu + \tilde{\nu}},$$

kur

$$\nu = tr(L),\ \tilde{\nu} = tr(L^T L) = \sum_{i=1}^n \parallel l(X_i) \parallel^2 .$$

Ja r ir pietiekami gluda, $\nu = o(n)$, un $\tilde{\nu} = o(n)$, tad $\hat{\sigma}^2$ ir nenovirzīts novērtējums parametram σ^2.

Teorēma 6. *[Rice (1984)] Ja r ir pietiekami gluda, tad*

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$$

ir nenovirzīts novērtējums, kur $X_1 \leq X_2 \ldots \leq X_n$.

Ja $\sigma(X_i) = \sigma < \infty$, t.i., nav atkarīga no prediktora, tad doto modeli sauc par homoskedastisku regresijas modeli, pretējā gadījumā - par heteroskedastisku regresijas modeli.

Funkcijas novērtējums $\hat{r}_n(X)$ nav jūtīgs pret heteroskedascitāti. Tomēr, pie ticamības joslu konstruēšanas regresijas funkcijai r ir jāņem vērā dispersijas mainīgums. Yu un Jones 2004. gadā ieviesa sekojošo procedūru mainīgās dispersijas novērtēšanai:
Definēsim jauno mainīgo $Z_i = ln(Y_i - r(X_i))^2$ un $\delta_i = ln(\epsilon_i^2)$. Tad $Z_i = ln(\sigma^2(X_i)) + \delta_i$. Tātad jānovērtē $ln(\sigma^2(X))$, regresējot logaritmētus kvadrātu atlikumus uz X.

1. Jānovērtē $r(x)$ ar neparametrisko metodi, iegūstot $\hat{r}(x)$.
2. Jāaprēķina $Z_i = ln(Y_i - \hat{r}_n(X_i))^2$.
3. Jāveic neparametrisko regresiju Z pret X, iegūstot $ln(\sigma^2(X))$, novērtējumu $\hat{q}(X)$. No tā seko, ka

$$\hat{\sigma}^2(X) = e^{\hat{q}(X)}.$$

3. Butstrapa metode

Butstrapa datu pārkārtošanas metode ir neparametriskā metode, kura aizvieto daudzus tradicionālus sadalījumu nosacījumus un asimptotisku rezultātu aprēķinus. Terminu "butstraps" ieviesa Efrons 1979. gadā.

Pieņemsim, ka $X_1, X_2, \ldots, X_n$ ir neatkarīgi un vienādi sadalīti ar $X_1 \sim F$ un $T = T(X_1, \ldots, X_n, F)$ ir kāda funkcija. Uzskatīsim, ka $X_1, \ldots, X_n$ labi reprezentē īsto populācijas sadalījumu. Butstrapa ideja ir ģenerēt (simulēt) daudz izlašu no dotās un aproksimēt statistikas T sadalījumu, t.i., izvēlēties no dotās izlases jaunas neatkarīgas un vienādi sadalītas izlases no empīriskās sadalījuma funkcijas $\hat{F}_n$. Visiem novērojumiem ir vienāda $1/n$ varbūtība tikt atlasītiem. To arī sauc par neparametrisko butstrapu.

$\{X_{11}^*, \ldots, X_{1n}^*\}, \{X_{21}^*, \ldots, X_{2n}^*\}, \ldots, \{X_{B1}^*, \ldots, X_{Bn}^*\}$ ir iegūtās butstrapa izlases, kur B apzīmē butstrapoto izlašu skaitu. T sadalījumu punktā t, t.i., $P_F(T \leq t)$, aproksimē ar $\{j$ skaits : $T_j^* \leq t\}/B$, kur $T_1^*, \ldots, T_B^*$ apzīmē statistikas T vērtības B dažādajām butstrapa izlasēm.

Ja butstrapotās izlases izvēlas no $F_{\hat{\theta}}$, kur $\hat{\theta}$ ir parametra θ novērtējums, to sauc par parametrisko butstrapu.

Butstrapa metodei ir savas priekšrocības. Sadalījumam nav nepieciešami īpaši nosacījumi (piemēram, lai kļūdas būtu sadalītas pēc Normālā sadalījuma), butstraps var nodrošināt precīzākus rezultātus gadījumā, kad dati slikti uzvedās vai izlases apjoms nav liels. Butstrapu var pielietot statistikai ar izlases sadalījumu, kuru ir grūti iegūt pat asimptotiski.

Regresijas gadījumā bieži vien lieto neparametrisko butstrapu. Bet, ja dati ir atkarīgi, nepieciešamas citas sarežģītākas butstrapa metodes (bloku stacionāro butstrapu utt.) Šī iemesla dēļ regresijas modeļos parasti tiek izmantots tā sauktais rezidiju butstraps, kas pārkārto atlikumus. Kā zināms, labam regresijas modelim atlikumi parasti ir neatkarīgi un vienādi sadalīti. Teorētisko pamatojumu rezidiju butstrapam deva Bickel un

Freedman 1981. gadā (skatīt [7]). Tomēr heteroskedastiskiem modeļiem neparametriskā butstrapa metode nav īsti piemērotāa (skatīt [8]). Tāpēc šiem modeļiem tika ieviests pāra un "mežonīgā" butstrapa metodes (skatīt [9],[10], [11],[2], [8]).

Šajā nodaļā butstrapa metodes apskatīsim heteroskedastiskā lineārās regresijas modeļa formā:

$$Y_i = r_i + \epsilon_i = \sum_{j=1}^{n} \beta_j X_{ij} + \epsilon_i, \;\; i = 1, \ldots, n,$$

kur $Y = (Y_1, \ldots, Y_n)^T$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$, $\beta = (\beta_1, \ldots, \beta_p)^T$, $\mathbf{X}$ ir $n \times p$ dizaina matrica, $E\epsilon_i = 0$, $D\epsilon_i = \sigma_i^2$. σ_i^2 ir nezināmā kļūdu dispersija. Apzīmēsim ar $\hat{r}_i = X_i^T \hat{\beta}$ neparametriskās regresijas funkcijas novērtējumu, tad $\hat{\epsilon}_i = Y_i - \hat{r}_i$ ir $i-$tais rezidijs.

Šajā nodaļā tiks apskatītas trīs butstrapa metodes regresijas modelī: rezidiju butstraps (residual bootstrap), pāru butstraps (pairs bootstrap) un mežonīgais butstraps (wild bootstrap).

Rezidiju butstraps.[8], [2], [9], [11], [6]

Tiek ģenerēti $\{\epsilon_i^*\}_1^n$, kas ir neatkarīgi un vienādi sadalīti, no normētu rezidiju izlases $\{\hat{\epsilon}_i/\sqrt{1 - pn^{-1}}\}_1^n$. Tādējādi butstrapa novērojumi tiek definēti kā $Y_i^* = \hat{r}_i + \epsilon_i^*$. Butstrapa mazāko kvadrātu novērtējums ir

$$\beta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y^* \tag{3.0.1}$$

Gadījumā, kad ir nelineārs novērtējums $\hat{\theta} = g(\hat{\beta})$, butstrapa dispersijas novērtējums ir $D(\beta^*) = E(\theta^* - \hat{\theta})(\theta^* - \hat{\theta})^T$, kur $\theta^* = g(\beta^*)$, β^* ir definēts kā (3.0.1). Ja $\theta = \beta$, tad $E\beta^* = \hat{\beta}$ un $D(\beta^*) = D(\hat{\beta}_{OLS}) = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$, kur $\hat{\sigma}^2 = (n-p)^{-1}\sum_{i=1}^{n} \hat{\epsilon}_i^2$. Tas nozīmē, ka butstrapa dispersijas novērtējums ir vienāds ar $\hat{\beta}_{OLS}$ dispersijas novērtējumu. Tāpēc, gadījumā, kad ir homoskedastisks regresijas modelis, t.i., $\sigma_i^2 = \sigma^2$, $D(\beta^*)$ ir nenovirzīts novērtējums. Bet, gadījumā, kad ir heteroskedastisks regresijas modelis, $D(\beta^*)$ parasti ir novirzīts novērtējums, jo

$$D(\hat{\beta}_{OLS}) = (\mathbf{X}^T\mathbf{X})^{-1}\sum_{i=1}^{n} \sigma_i^2 X_i X_i^T (\mathbf{X}^T\mathbf{X})^{-1}.$$

Gadījumā, ja ir rezidiju nehomogenitāte, šī informācija pazūd datu ģenerēšanas procesā un vairs neietekmē novērtējumu $D(\beta^*)$.

Pāru butstraps. [2], [9], [11], [6]

Tiek ģenerēti neatkarīgi un vienādi sadalīti $\{Y_i^*, X_i^*\}_1^n$ no $\{(Y_i, X_i)\}_1^n$. Butstrapa mazāko kvadrātu novērtējums ir

$$\beta^* = \left(\sum_{i=1}^n X_i^* X_i^{*T}\right)^{-1} \sum_{i=1}^n X_i^* Y_i^* \tag{3.0.2}$$

un attiecīgais butstrapa dispersijas novērtējums ir $D(\beta^*) = E(\theta^* - \hat{\theta})(\theta^* - \hat{\theta})^T$, kur $\hat{\theta} = g(\hat{\beta})$, $\theta^* = g(\beta^*)$, β^* ir definēts kā (3.0.2).

Šis metodes trūkums ir tāds, ka tiek ignorēta datu $\{Y_i, X_i\}$ nehomogenitāte.

Mežonīgais butstraps. [2], [9], [11], [10], [6]

Iepriekš apskatītā rezidiju butstrapa metode nestrādā labi heteroskedastiskajam regresijas modelim. Tāpēc šeit tiek apskatīts cits rezidiju butstrapošanas paņēmiens, tā ir mežonīgā butstrapa metode. Šo metodi attīstīja R. Y. Liu 1988. gadā, sekojot C. F. J. Wu un R. Beran darbiem.

Butstrapoti dati ir formā

$$Y_i^* = X_i^T \hat{\beta} + a_i \hat{\epsilon}_i t_i^*, \;\; i = 1, \ldots, n,$$

kur $a_i = 1/\sqrt{1 - w_i}$, $w_i = X_i^T (\mathbf{X}^T \mathbf{X})^{-1} X_i$ un $t^* = (t_i^*)_1^n$ tiek iegūts pēc datu pārkārtošanas metodes ar nosacījumu, ka

$$Et^* = 0 \;\; \text{un} \;\; D(t^*) = I. \tag{3.0.3}$$

Butstrapa mazāko kvadrātu novērtējums, kas balstās uz Y_i^*, ir

$$\hat{\beta}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T Y^*, \;\; \text{kur} \;\; Y^* = (Y_1^*, \ldots, Y_n^*).$$

Secinājumus par β vai $\theta = g(\beta)$ var iegūt no izkliedes ap $\hat{\beta}^*$. Piemēram, $D(\hat{\theta})$ var novērtēt ar $D(\hat{\beta}^*) = E(\hat{\theta}^* - \hat{\theta})(\hat{\theta}^* - \hat{\theta})^T$, kur $\hat{\theta}^* = g(\hat{\beta}^*)$, $\hat{\theta} = g(\hat{\beta})$. Ja parametrs β ir lineārs, tad $E(\hat{\beta}^*) = \hat{\beta}$ un

$$D(\hat{\beta}^*) = (\mathbf{X}^T \mathbf{X})^{-1} \sum_{i=1}^n a_i^2 \hat{\epsilon}_i^2 X_i X_i^T (\mathbf{X}^T \mathbf{X})^{-1}.$$

Eksistē vairākās datu pārkārtošanas metodes, kas apmierina nosacījumu (3.0.3). Savā darbā C. F. J. Wu (1986) [2] aprakstīja divas tādas datu pārkārtošanas metodes:

1. Kļūdas t_i^* tiek atlasītas no Hadamarda matricas (Hadamard matrix) $[\delta_i^{(k)}]$, $\delta_i^{(k)} = \pm 1$, $1 \le i \le n$, $1 \le k \le R$, kura apmierina nosacījumus, ka

$$\sum_{k=1}^R \delta_i^{(k)} = 0 \;\; \text{visiem} \;\; i,$$

$$\frac{1}{R}\sum_{k=1}^{R}\delta_i^{(k)}\delta_j^{(k)} = 0, \text{ ja } i \neq j.$$

Parasti $n+1 \leq R \leq n+4$. Butstrapotās vērtības $Y^{(k)} = (Y_i^{(k)})_{i=1}^n$ tiek definētas kā

$$Y_i^{(k)} = X_i^T\hat{\beta} + a_i\hat{\epsilon}_i\delta_i^k, \;\; i = 1, \ldots, n, \tag{3.0.4}$$

kur $a_i = 1/\sqrt{1-w_i}$. Butstrapa mazāko kvadrātu novērtējums ir

$$\hat{\beta}^{(k)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y^{(k)},$$

bet dispersijas novērtējums ir

$$D(\hat{\beta}^{(k)}) = \frac{1}{R}\sum_{k=1}^{R}(\hat{\theta}^{(k)} - \hat{\theta})(\hat{\theta}^{(k)} - \hat{\theta})^T, \;\; \hat{\theta}^{(k)} = g(\hat{\beta}^{(k)}).$$

Šeit $D(\hat{\beta}^{(k)})$ ir nenovirzīts novērtējums. Dotā metode tiek domāta tādiem gadījumiem, kad kļūdas ϵ_i ir simetriskas, jo katram rezidijam $\hat{\epsilon}_i$ puse no R butstrapa izlasēm satur $+\hat{\epsilon}_i$, bet otrā puse satur $-\hat{\epsilon}_i$.

2. Kļūdas $\{t_i^*\}_1^n$ ir butstrapa izlase no ierobežotās populācijas $\{c_j\}_{j=1}^M$ ar

$$\sum_{j=1}^{M} c_j = 0, \qquad \frac{1}{M}\sum_{j=1}^{M} c_j^2 = 1.$$

Viens no veidiem, kā var izvēlēties $\{c_j\}$ ir

$$c_j = (\hat{\epsilon}_j - \bar{\hat{\epsilon}}) \Bigg/ \left[\frac{1}{n}\sum_{j=1}^{n}(\hat{\epsilon}_j - \bar{\hat{\epsilon}})^2\right]^{1/2}, \;\; j = 1, \ldots, n.$$

Mammens (1993) [9] uzskata, ka viens no populārākajiem veidiem kā var izvēlēties t^* sadalījumu ir divu punktu sadalījums

$$F_1: \;\; t_i^* = \begin{cases} -(\sqrt{5}-1)/2 & \text{ar varbūtību } \; p = (\sqrt{5}+1)/2\sqrt{5}, \\ (\sqrt{5}+1)/2 & \text{ar varbūtību } \; 1-p. \end{cases}$$

Davidsons un Flachaire (2001) [10] parādīja, ka Radamacher sadalījums

$$F_2: \;\; t_i^* = \begin{cases} 1 & \text{ar varbūtību } \; 0,5 \\ -1 & \text{ar varbūtību } \; 0,5. \end{cases}$$

vienmēr dod labākus rezultātus nekā F_1 sadalījums.

Dažos literatūras avotos palīgfunkcijas $a_i = 1/\sqrt{1-w_i}$ vietā izmanto arī citas funkcijas:

$$a_i^{(1)} = 1, \quad a_i^{(2)} = \sqrt{\frac{n}{n-p}}, \quad a_i^{(3)} = \frac{1}{1-w_i}.$$

Eksperimentu rezultātā Davidsons un Flachaire (2001) ieguva, ka rezidiju transformācija $a_i^{(3)}$ ir labāka nekā pārējās trīs versijas. Līdzīgi rezultāti tika iegūti Flachaire (1999), Longa un Ervina (2000), Chestera un Jewitta(1987) un citu pētnieku darbos.

Mežonīgais butstraps salīdzinājumā ar pāru butstrapu. [9], [2]

Salīdzināsim mežonīgo butstrapu ar pāru butstrapu regresijas modelī. Pāru butstrapa regresori tiek ģenerēti pēc datu pārkārtošanas metodes ar atkārtojumiem. Tātad šie regresori nav atkarīgie mainīgie. Tā kā mēs ģenerējam regresorus un rezidijus vienlaicīgi, tad $E(\epsilon_i^*|X_i^*) \neq 0$. Īstajā datu ģenerēšanas procesā tiek pieņemts, ka regresori X ir neatkarīgi un $E(\epsilon_i|X_i) = 0$. Tādējādi datu ģenerēšanas procesā ar pāru butstrapu šie nosacījumi neizpildās. Butstrapa metodes principi paredz, ka butstrapotu datu ģenerēšanas process būs pēc iespējas tuvāk īstajam datu ģenerēšanas procesam. Mūsu gadījumā var uzlabot pāru butstrapa efektivitāti ar sekojošas modifikācijas palīdzību:

1. Ja mēs ģenerēsim (X^*, ϵ^*) no $(X, \hat{\epsilon})$, tad nosacījums, ka $E(\epsilon_i^*|X_i^*) = 0$, izpildīsies. Un attiecīgais butstrapotu datu ģenerējošais process būs

$$Y_i^* = X_i^*\hat{\beta} + \epsilon_i^* t_i^*,$$

kur t_i^* ir savstarpēji neatkarīgas izlases no tāda palīgsadalījuma, ka $E(t_i^*) = 0$ un $E(t_i^{*2}) = 1$.

2. Ja regresorus ģenerēt, neizmantojot datu pārkārtošanas metodi, tad X^* būs neatkarīgi. Nevar ģenerēt rezidijus ar datu pārkārtošanas metodi neatkarīgi no regresoriem, jo heteroskedascitāte var būt funkcija no tiem pašiem regresoriem. Tādā veidā iegūstam sekojošu butstrapotu datu ģenerēšanas procesu

$$Y_i^* = X_i\hat{\beta} + \hat{\epsilon}_i t_i^*.$$

Tātad pēc divu nosacījumu ieviešanas pāru butstrapa datu ģenerēšanas procesā, lai tas būtu tuvāk īstajam datu ģenerēšanas procesam, mēs iegūstam mežonīga butstrapa datu ģenerēšanas procesu.

4. Vienlaicīgo ticamības joslu konstruēšana neparametriskajai regresijai

Vienlaicīgās ticamības joslas formā

$$\hat{r}(x) \pm w(x)$$

funkcijas $r(x)$ novērtējumam ar pārklājuma varbūtību $1-\alpha$ apmierina nosacījumu, ka

$$P\left(\hat{r}(x) - w(x) \leq r(x) \leq \hat{r}(x) + w(x) \ \ \forall x\right) = 1 - \alpha$$

.

4.1. San un Loader metode

Pieņemsim, ka apskatāmais modelis ir formā

$$Y_i = r(X_i) + \epsilon_i, \quad ,$$

kur ϵ_i ir neatkarīgi un sadalīti pēc Normālā sadalījuma ar vidējo vērtību 0 un dispersiju σ^2. Šajā nodaļā mēs apskatīsim ticamības joslu konstruēšanas regresijas funkcijai $r(x)$ klasisko neparametrisko pieeju. Parasti šīs joslas ir formā

$$\hat{r}_n(x) \pm c\ se(x), \tag{4.1.1}$$

kur $se(x)$ ir funkcijas $\hat{r}_n(x)$ standartnovirzes novērtējums un $c > 0$ ir konstante.

Biasa problēma.

Parasti ticamības joslas veidā (4.1.1) tiek konstruētas funkcijai $\bar{r}_n(x) = E(\hat{r}_n(x))$ nevis funkcijai $r(x)$. Ticamības intervālu konstruēšana īstai regresijas funkcijai ir diezgan komplicēts darbs, tam iemesls tiek paskaidrots zemāk.

Apzīmēsim $\hat{r}_n(x)$ vidējo vērtību ar $\bar{r}(x)$, bet standartnovirzi ar $s_n(x)$. Tad

$$\frac{\hat{r}_n(x) - r(x)}{s_n(x)} = \frac{\hat{r}_n(x) - \bar{r}_n(x)}{s_n(x)} + \frac{\bar{r}_n(x) - r(x)}{s_n(x)}$$

$$= Z_n(x) + \frac{bias(\hat{r}_n(x))}{\sqrt{D(\hat{r}_n(x))}},$$

kur $Z_n(X) = (\hat{r}_n(x) - \bar{r}_n(x))/s_n(x)$. Pirmais saskaitāmais $Z_n(x)$ konverģē uz standartnormālu sadalījumu, bet otrais - uz 0. Otrais saskaitāmais paliks pat pie lieliem izlases apjomiem. Rezultātā ticamības joslas nebūs centrētas apkārt īstai funkcijai r.

Pieņemsim, ka $\hat{r}_n(x)$ ir lineārs gludinātājs, tā ka $\hat{r}_n(x) = \sum_{i=1}^n Y_i l_i(x)$. Tad

$$\bar{r}(x) = E(\hat{r}_n(x)) = \sum_{i=1}^n l_i(X) r(X_i).$$

Tātad apskatīsim divus gadījumus, kad ir homoskedastisks un heteroskedastisks modelis.

1. $\sigma^2(X) = \sigma^2 = D(\epsilon_i)$ ir konstanta. Tad $D(\hat{r}_n(x)) = \sigma^2 \|l(x)\|^2$. Apskatīsim ticamības joslas funkcijai $\bar{r}_n(x)$ formā

   $$I_{SL}(x) = (\hat{r}_n(x) - c\hat{\sigma}\|l(x)\|,\ \hat{r}_n(x) + c\hat{\sigma}\|l(x)\|)$$

   kādam $c > 0$ un $a \le x \le b$. Sekojot San un Loader (1994)[1] pieejai, sākumā pieņemsim, ka σ ir zināma. Tad

   $$P\left(\bar{r}(x) \notin I(x),\ \ x \in [a,b]\right) = P\left(\max_{x\in[a,b]} \frac{\mid \hat{r}(x) - \bar{r}(x) \mid}{\sigma\|l(x)\|} > c\right)$$

   $$= P\left(\max_{x\in[a,b]} \frac{\mid \sum_i \epsilon_i l_i(x) \mid}{\sigma\|l(x)\|} > c\right) = P\left(\max_{x\in[a,b]} \mid W(x) \mid > c\right),$$

   kur $W(x) = \sum_{i=1}^n z_i T_i(x)$, $z_i = \epsilon_i/\sigma \sim N(0,1)$ un $T_i(x) = l_i(x)/\|l(x)\|$. $W(x)$ ir Gausa process. Lai atrastu c, ir nepieciešams noteikt Gausa procesa maksimuma sadalījumu. San un Loader (1994)[1] parādīja, ka

   $$P\left(\max_x \left|\sum_{i=1}^n z_i T_i(x)\right| > c\right) \approx 2(1 - \Phi(c)) + \frac{k_0}{\pi} e^{-c^2/2}. \qquad (4.1.2)$$

   Šī formula ir spēkā lieliem c, kur

   $$k_0 = \int_a^b \|T'(x)\| dx,$$

   $T'(X) = (T_1'(X), \ldots, T_n'(X))$ un $T_i'(X) = \partial T_i(X)/\partial X$. Vienādojumu (4.1.2) angliski sauc par "tube" formulu. Šis formulas pamatojums ir sekojošs.

□ Pieņemsim, ka $W(x) = \sum_{i=1}^{n} z_i T_i(x)$ un $\|T(x)\| = \sum_{i=1}^{n} T_i(x) = 1$, vektors $T(x)$ ir uz vienības sfēras katrā punktā x. Tā kā $z = (z_1, \ldots, z_n)$ tiek sadalīts pēc daudzdimensiju Normālā sadalījuma, tad

$$P\left(\sup_x W(x) > c\right) = P\left(\sup_x \langle z, T(x)\rangle > c\right) =$$

$$= P\left(\sup_x \left\langle \frac{z}{\|z\|}, T(x)\right\rangle > \frac{c}{\|z\|}\right) =$$

$$= \int_{c^2}^{\infty} P\left(\sup_x \langle U, T(X)\rangle > \frac{c}{\sqrt{y}}\right) h(y)dy, \tag{4.1.3}$$

kur $U = (U_1, \ldots, U_n)$ ir vienmērīgi sadalīts uz $n-1-$dimensionālās vienības sfēras S un $h(y)$ ir ar n brīvības pakāpēm χ^2 blīvuma funkcija. Tā kā $\|U - T(x)\|^2 = 2(1 - \langle U, T(x)\rangle)$, var redzēt, ka $\sup_x \langle U, T(x)\rangle > \frac{c}{\sqrt{y}}$ tad un tikai tad, ja $U \in tube(r, M)$, kur $r = \sqrt{2(1 - c/\sqrt{y})}$, $M = \{T(X)\colon x \in \mathbf{x}\}$ ir kopa uz sfēras S,

$$tube(r, M) = \{u\colon d(u, M) \leqslant r\}$$

un

$$d(u, M) = \inf_{x \in \mathbf{x}} \|u - T(x)\|.$$

No tā seko, ka

$$P\left(\sup_x \langle U, T(x)\rangle > \frac{c}{\sqrt{y}}\right) = P(U \in tube(r, M)).$$

$A_n = 2\pi^{n/2}/\Gamma(n/2)$ ir vienības sfēras laukums. Tātad

$$P(U \in tube(r, M)) = \frac{volume(tube(r, M))}{A_n}.$$

Naimans (1990) [12] izveda kopas $tube(r, M)$ apjoma formulu

$$k_0 \frac{A_n}{A_2} P\left(B_{1,(n-2)/2} \geq \frac{c^2}{y}\right) + \frac{A_n}{A_1} P\left(B_{1/2,(n-1)/2} \geq \frac{c^2}{y}\right),$$

kur $B_{\varepsilon,r}$ ir vaļējā lode ar centru ε un rādiusu r. Ja ievietosim šo formulu integrālā (4.1.3) un neievērosim tos izteiksmes locekļus, kas ir mazāki par $c^{-1/2}e^{-c^2/2}$, tad iegūsim formulu (4.1.2). ■

Piezīme 2. *Lekciju kursā netika apskatīts konstantes k_0 aprēķins, jo tas nav triviāls uzdevums. Šajā darbā dotā konstante tiek aprēķināta pēc metodes, kas tiek aprakstīta Faraway un San (1995)[13] un San, Raz un Faraway (1999) darbos [14].*

Tika piedāvāta k_0 aproksimācija, gadījumā, ja $X = [a, b]$. Sadalot intervālu $[a, b]$ m punktos $a = t_0 < \ldots < t_m = b$, iegūstam

$$k_0 = \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \|T'(x)\| dx \approx \sum_{i=1}^{m} \|T(t_i) - T(t_{i-1})\|. \tag{4.1.4}$$

Ja c tiek aprēķināts no vienādojuma

$$2(1 - \Phi(c)) + \frac{k_0}{\pi} e^{-c^2/2} = \alpha, \tag{4.1.5}$$

tad iegūstam prasītās vienlaicīgās ticamības joslas.

Ja σ ir nezināma, tad lieto novērtējumu $\hat{\sigma}$.

San un Loader [1] piedāvā aizvietot vienādojuma (4.1.2) labo pusi ar

$$P(\mid T_m \mid > c) + \frac{k_0}{\pi} \left(1 + \frac{c^2}{m}\right)^{-m/2},$$

kur T_m ir sadalīts pēc $t-$sadalījuma ar $m = n - tr(L)$ brīvības pakāpēm.

2. $\sigma^2(x)$ ir funkcija no x. Tad $D(\hat{r}_n(x)) = \sum_{i=1}^{n} \sigma^2(X_i) l_i^2(x)$. Šajā gadījumā ticamības joslas ir formā

$$I_{SL}(x) = \hat{r}_n(x) \pm cs(x),$$

kur

$$s(x) = \sqrt{\sum_{i=1}^{n} \hat{\sigma}^2(X_i) l_i^2(x)}.$$

$\hat{\sigma}(x)$ ir $\sigma(x)$ novērtējums un c ir konstanta, kas tiek rēķināta pēc formulas (4.1.5). Ja $\hat{\sigma}(x)$ mainās lēni pēc x, tad $\sigma(X_i) \approx \sigma(x)$ tādiem i, kuriem $l_i(x)$ ir liels, tāpēc $s(x) \approx \hat{\sigma}(x)\|l(x)\|$. Tādējādi aptuvenās ticamības joslas ir

$$I_{SL}(x) = \hat{r}_n(x) \pm c\hat{\sigma}(x)\|l(x)\|.$$

4.2. Mežonīgā butstrapa metode

Šajā nodaļā tiek apskatītas metodes asimptotisko vienlaicīgo ticamības joslu konstruēšanai neparametriskajai regresijai. Apskatāmais modelis ir heteroskedastisks un novērojumi nav vienādi sadalīti. Vienlaicīgu ticamības joslu konstruēšana balstās uz lokālo polinomu novērtējumu, bet attiecīgā kvantile tiek iegūta ar mežonīgā butstrapa metodi.

Pieņemsim, ka apskatāmais modelis ir formā

$$Y_i = r(X_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

kur kļūdas ϵ_i ir neatkarīgas, bet neobligāti vienmērīgi sadalītas, ar $E\epsilon_i = 0$, $D\epsilon_i = \sigma^2(X_i)$, kas apmierina nosacījumu

$$0 < \sigma^2_{inf} \leq \sigma^2(X_i) \leq \sigma^2_{sup} < \infty, \;\; E|\epsilon_i|^M \leq C(M) < \infty \;\; \text{visiem } i \text{ un } M. \tag{4.2.1}$$

Ja $\mathbf{X}$ nav gadījuma dizains, tad uzskatām, ka eksistē konstantes $0 < C_1 \leq C_2 < \infty$ tādas, ka

$$C_1(n(b-a) - \log n) \leq \#\{i | X_i \in [a,b)\}$$

$$\leq C_2(n(b-a) + \log n) \;\; \text{visiem} \;\; 0 \leq a < b \leq 1. \tag{4.2.2}$$

Pieņemsim, ka

$$r(x) \in C[0,1]. \tag{4.2.3}$$

$\hat{r}(x)$ ir $p-$tās kārtas lokālais polinomu novērtējums funkcijai $r(x)$. Šajā darbā mēs apskatām vienlaicīgās ticamības joslas neparametriskai regresijai formā

$$I_x = [\hat{r}(x) - t(x), \hat{r}(x) + t(x)].$$

$t(x)$ vērtības tiek definētas tādā veidā, lai

$$P(r(x) \in I_x \;\; \text{visiem} \;\; X \in [0,1]) \longrightarrow 1 - \alpha$$

kādam uzdotam nozīmības līmenim α, $0 < \alpha < 1$.

Speciālajā gadījumā, kad ϵ_i ir neatkarīgi un vienādi sadalīti, Eubanks un Speckmans (1993)[15] aproksimēja procesu $\{(\hat{r}(x) - r(x))/\sqrt{D(\hat{r}(x))}\}_{x \in [0,1]}$ ar kādu stacionāru Gausa procesu un noteica procesa absolūto vērtību maksimuma asimptotisku $(1-\alpha)-$kvantili, balstoties uz Bickela un Rosenblatta (1973)[16] rezultātiem. Situācijā, kad pastāv heteroskedascitāte, nevar rīkoties līdzīgi. Tāpēc izmantosim butstrapa metodes parasto ideju. Izlaižot kāda Gausa procesa maksimālas novirzes sadalījuma aproksimācijas soli, mēs gribam sagaidīt labāku pārklājuma precizitāti vienlaicīgajām ticamības joslām. Heteroskedastisko kļūdu dēļ mēs izmantosim mežonīgu butstrapu un neparametriskās regresijas lokālo polinomu novērtējumu.

Tātad no sākuma iegūstam rezidijus $\hat{\epsilon}_i = Y_i - \hat{r}(X_i)$. Tālāk ģenerējam neatkarīgus gadījuma lielumus ϵ^*_i ar vidējo vērtību 0, dispersiju $\hat{\epsilon}^2_i$ un ierobežotiem augstākās kārtas

momentiem. Pielietosim mežonīgā butstrapa vienkāršoto variantu

(i) $\epsilon_i^* \sim N(0, \hat{\epsilon}_i^2)$

vai

(ii) $P(\epsilon_i^* = -\hat{\epsilon}_i) = P(\epsilon_i^* = +\hat{\epsilon}_i) = \frac{1}{2}$.

Nākamais solis ir procesa $\{\hat{r}(x) - r(x)\}_{x\in[0,1]}$ stohastiskās daļas $\hat{r}_0(x) = \sum l_i(x)\epsilon_i$ simulēšana ar $\hat{r}_0^*(x) = \sum l_i(x)\epsilon_i^*$.

Ja mēs salīdzināsim divu gadījuma lielumu kumulatīvās sadalījuma funkcijas, tad mēs varam gaidīt, ka tās būs tuvas viena otrai, ja ar lielu varbūtību atšķirības starp gadījuma lielumiem ir mazas. Sakarā ar to definēsim sekojošo jēdzienu.

Definīcija 12. *Pieņemsim, ka* $\{Y_n\}$ *un* $\{Z_n\}$ ($Z_n \geq 0$ *gandrīz droši*) *ir gadījuma lieluma virknes, un* $\{\gamma_n\}$ *ir pozitīvo reālo skaitļu virkne. Ja* $P(|Y_n| > CZ_n) \leq C\gamma_n$ *ir spēkā pie* $n \geq 1$ *un* $C < \infty$, *tad raksta*

$$Y_n = \widetilde{O}(Z_n, \gamma_n).$$

Sekojošā lemma parāda kādā veidā $\widetilde{O}$ var izmantot, lai pierādītu divu gadījuma lielumu tuvumu.

Lemma 1. *[17] Pieņemsim, ka* $\{X_n\}$ *ir gadījuma lielumu virkne ar* $P(X_n \in [a,b]) \leq C((b-a)c_n + \gamma_{n1})$ *pie patvaļīgiem* a *un* b *un* $Y_n = \widetilde{O}(\gamma_{n2}, \gamma_{n3})$. *Tad*

$$P(X_n + Y_n < t) = P(X_n < t) + O(\gamma_{n1} + c_n\gamma_{n2} + \gamma_{n3})$$

uzvedas vienmērīgi pie $t \in (-\infty, \infty)$.

Tātad tika iegūts svarīgs rezultāts, kas balstās uz procesa $\{\hat{r}(x) - r(x)\}_{x\in[0,1]}$ aproksimāciju ar $\{\hat{r}_0^*(x)\}_{x\in[0,1]}$. Tādējādi tika sasaistīti nosacīti sadalījumi $\mathcal{L}(\epsilon_i^*|Y)$ ar sadalījumiem $\mathcal{L}(\epsilon_i)$. Turpmāk ar $\delta > 0$ sapratīsim patvaļīgu ļoti mazu konstanti, bet ar $\lambda < \infty-$patvaļīgu ļoti lielu konstanti.

Teorēma 7. *[17] Pieņemsim, ka izpildās nosacījumi* (4.2.1), (4.2.2), (4.2.3) *un* ϵ_i^*, $i = 1, \ldots, n$ *tiek definēti pēc iepriekš aprakstītās procedūras. Tad eksistē* $\{\epsilon_i\}$ *un* $\{\epsilon_i^*\}$ *versijas uz attiecīgās kopējās varbūtību telpas tādas, ka*

$$\sup_{x\in[0,1]} \{|(\hat{r}(x) - r(x)) - \hat{r}_0^*(x)|\} = \widetilde{O}(n^{\delta}(nh)^{-1} + h^k, n^{-\lambda}),$$

kas ir spēkā uz kopas $(\epsilon_1, \ldots, \epsilon_n) \in \Omega_0$ *ar* $P(\Omega_0) \geq 1 - O(n^{-\lambda})$.

Vienlaicīgās ticamības joslas ar vienmērīgo platumu.

Pieņemsim, ka t^*_α ir lieluma

$$U^*_{n0} = \sup_{x\in[0,1]} \{|\hat{r}^*_0(x)|\}$$

sadalījuma $(1-\alpha)-$kvantile. Šis lielums tika ieviests, lai simulētu

$$U_n = \sup_{x\in[0,1]} \{|\hat{r}(x) - r(x)|\}.$$

No Teorēmas 7 ir zināms, ka process $\hat{r}(x) - r(x)$ ir līdzīgs nosacītam procesam $\hat{r}^*_0(x)$ uz attiecīgās varbūtību telpas. Nākamā lemma definē apakšējo robežu varbūtībai, ka $\sup_x\{|\hat{r}^*_0(x)|\}$ iekrīt mazos intervālos.

Lemma 2. *[17] Pieņemsim, ka izpildās nosacījumi* (4.2.1), (4.2.2), (4.2.3). *Tad*

$$P\left(\sup_{x\in[0,1]} \{|\hat{r}^*_0(x)|\} \in [a,b]\right) = O((b-a)(nh)^{1/2}(\log n)^{1/2} + n^\delta; (nh)^{-1/2}).$$

Nākamā teorēma definē kļūdas augšējo robežu pārklājuma precizitātei vienlaicīgām ticamības joslām ar vienmērīgu platumu t^*_α apkārt $\hat{r}(x)$.

Teorēma 8. *Pieņemsim, ka izpildās nosacījumi* (4.2.1), (4.2.2), (4.2.3). *Tad*

$$P(r(x) \in [\hat{r}(x) - t^*_\alpha, \hat{r}(x) + t^*_\alpha] \quad visiem \quad x \in [0,1]) =$$

$$= 1 - \alpha + O(n^\delta (nh)^{-1/2} + (nh)^{1/2}(\log n)^{1/2} h^k).$$

□ Pēc Teorēmas 7 iegūstam, ka

$$|U_n - U^*_{n0}| \le \sup_{x\in[0,1]} \{|(\hat{r}(x) - r(x)) - \hat{r}^*_0(x)|\} =$$

$$= \widetilde{O}(n^\delta (nh)^{-1} + h^k, n^{-\lambda-1})$$

ir spēkā attiecīgā varbūtību telpā, kas nodrošina (pēc Lemmām 1 un 2)

$$\sup_t \{|P(U_n < t) - P(U^*_{n0} < t|Y)|\} =$$

$$= \widetilde{O}(n^\delta (nh)^{-1/2} + h^k (nh)^{1/2}(\log n)^{1/2}, n^{-\lambda})$$

vienmērīgi noteiktā kopā $Y \in \Omega_0$ ar $P(\overline{\Omega}_0) = O(n^{-\lambda})$. Tas nozīmē, ka

$$P(U_n < t)|_{t=t^*_\alpha} = P(U^*_{n0} < t^*_\alpha|Y) + \widetilde{O}(n^\delta (nh)^{-1/2} +$$

$$+h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda}) =$$

$$= 1 - \alpha + \widetilde{O}(n^{\delta}(nh)^{-1/2} +$$

$$+h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda}),$$

telpā $Y \in \Omega_0$. Integrējot pa t^*_α, iegūstam izteiksmi, kuru vajadzēja pierādīt. ■

Vienlaicīgās ticamības joslas ar mainīgo platumu.

Šeit mēs piedāvāsim lietderīgo alternatīvu iepriekš aprakstītajām vienlaicīgām ticamības joslām. Ideja ir tāda, ka joslu platums ir proporcionāls $\hat{r}(x)$ novērtētai standartnovirzei. Rezidiji $\hat{\epsilon}_i$ tiek izmantoti, lai novērtētu $\sigma^2(x) = D(\hat{r}(x))$ ar

$$\hat{\sigma}^2(x) = \sum_i l_i^2(x)\hat{\epsilon}_i^2.$$

Pieņemsim, ka t^{**}_α ir lieluma

$$T^*_{n0} = \sup_{x\in[0,1]} \{|\hat{r}^*_0(x)| \ / \sqrt{\hat{\sigma}^2(x)}\}$$

sadalījuma $(1-\alpha)-$kvantile. Šis lielums simulē

$$T_n = \sup_{x\in[0,1]} \{|\hat{r}(x) - r(x)|/\sqrt{\hat{\sigma}^2(x)}\}.$$

Lemma 3. *Pieņemsim, ka izpildās nosacījumi* (4.2.1), (4.2.2), (4.2.3). *Tad*

$$\sup_{x\in[0,1]} \{|\hat{\sigma}^2(x) - \sigma^2(x)|\} = \widetilde{O}(n^{\delta}(nh)^{-3/2}, n^{-\lambda}).$$

□ $\hat{\epsilon}_i = Y_i - \hat{r}(X_i) = r(X_i) + \epsilon_i - \hat{r}(X_i)$, tad

$$\hat{\epsilon}_i^2 = \epsilon_i^2 - 2\epsilon_i(\hat{r}(X_i) - r(X_i)) + (\hat{r}(X_i) - r(X_i))^2.$$

No tā seko, ka pie fiksētā punkta x

$$\hat{\sigma}^2(x) - \sigma^2(x) \le \left|\sum_i l_i^2(x)[\epsilon_i^2 - \sigma_i^2]\right| +$$

$$+ \left|\sum_i l_i^2(x)[(\hat{r}(X_i) - r(X_i))^2 - 2\epsilon_i(\hat{r}(X_i) - r(X_i))]\right| =$$

$$= \widetilde{O}(n^{\delta}(nh)^{-3/2}, n^{-\lambda}).$$

Veicot procesa $\hat{\sigma}^2(x) - \sigma^2(x)$ aproksimāciju pietiekami smalki sašķeltā intervālā $[0,1]$, mēs pierādījām lemmas apgalvojumu. ■

No šis lemmas seko, ka $|\hat{r}(x) - r(x)|/\sqrt{\hat{\sigma}^2(x)}$ un $\hat{r}^*_0(x)/\sqrt{\hat{\sigma}^2(x)}$ var labi aproksimēt attiecīgi ar $|\hat{r}(x) - r(x)|/\sqrt{\sigma^2(x)}$ un $\hat{r}^*_0(x)/\sqrt{\sigma^2(x)}$.

Teorēma 9. *Pieņemsim, ka izpildās nosacījumi* (4.2.1), (4.2.2), (4.2.3). *Tad*

$$P\left(r(x) \in [\hat{r}(x) - \sqrt{\hat{\sigma}^2(x)}t_\alpha^{**}, \hat{r}(x) + \sqrt{\hat{\sigma}^2(x)}t_\alpha^{**}] \quad visiem \quad x \in [0,1]\right) =$$

$$= 1 - \alpha + O\left(n^\delta (nh)^{-1/2} + (nh)^{1/2}(\log n)^{1/2} h^k\right).$$

Teorēmas pierādījuma ideja ir tāda, ka, izmantojot dotās aproksimācijas

$$\frac{\hat{r}(x) - r(x)}{\sqrt{\hat{\sigma}^2(x)}} = \frac{\hat{r}(x) - r(x)}{\sqrt{\sigma^2(x)}} + \widetilde{O}(n^\delta (nh)^{-1}, n^{-\lambda})$$

un

$$\frac{\hat{r}_0^*(x)}{\sqrt{\hat{\sigma}^2(x)}} = \frac{\hat{r}_0^*(x)}{\sqrt{\sigma^2(x)}} + \widetilde{O}(n^\delta (nh)^{-1}, n^{-\lambda}),$$

var pierādīt teorēmas apgalvojumu līdzīgi kā Teorēmā 8 (skatīt [17]).

5. Simulācijas

Šajā nodaļā, izmantojot pārklājuma precizitāti, salīdzināsim mežonīgā butstrapa vienlaicīgās ticamības joslas, kas tiek konstruētas ar dažādu metožu palīdzību. Simulācijas balstās uz fiksēta dizaina modeli

$$Y_i = r(i/n) + \epsilon_i, \;\; E(\epsilon_i) = 0, \;\; D(\epsilon_i) = \sigma_i^2, \;\; i = 1, \ldots, n,$$

ar regresijas funkciju

$$r(x) = \exp(-32(x - 0.5)^2).$$

Tika ģenerēti dati pie izlašu apjomiem $n = 100$, $n = 200$ un $n = 300$, un tika apskatītas dažādas kļūdu dispersijas struktūras: homogēna dispersija $\sigma_i = 0.1$, vāja heteroskedascitāte $\sigma_i = 0.05 + 0.1r(X_i)$ un stipra heteroskedascitāte $\sigma_i = 0.01 + 0.2r(X_i)$. Kļūdas ϵ_i tiek simulētas pēc normāla sadalījuma ar vidējo vērtību 0 un dispersiju σ_i^2. Ģenerēto datu vienas simulācijas piemērs tiek atspoguļots Attēlos 5.1. un 5.2. Tiek salīdzinātas četras metodes vienlaicīgo ticamības joslu konstruēšanai neparametriskajai regresijai:

1. vienādā platuma ticamības joslas

$$I_x^U = [\hat{r}(x) - t_\alpha^*, \hat{r}(x) + t_\alpha^*],$$

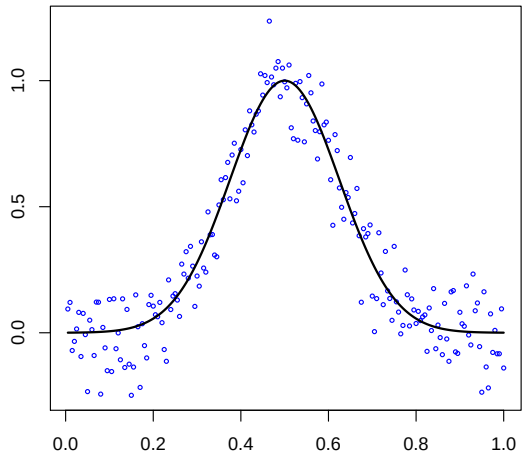
2. mainīgā platuma ticamības joslas

$$I_x^T = [\hat{r}(x) - \sqrt{\hat{\sigma}^2(x)} t_\alpha^{**}, \hat{r}(x) + \sqrt{\hat{\sigma}^2(x)} t_\alpha^{**}],$$

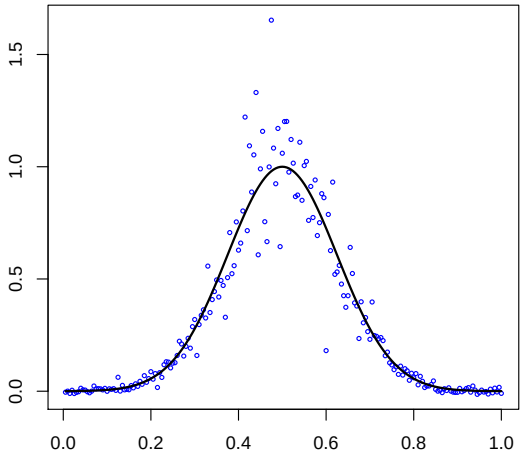
3. I_x^U modifikācija

$$I_x^W = [\hat{r}(x) - \sqrt{w(x)} t_\alpha^{W*}, \hat{r}(x) + \sqrt{w(x)} t_\alpha^{W*}],$$

kur t_α^{W*} ir lieluma $W_{n0}^* = \sup_{x \in [0,1]} \{|\hat{r}_0^*(x)|/\sqrt{w(x)}\}$ sadalījuma $(1-\alpha)-$kvantile un $w(x) = \sum_j l_j^2(x)$ ir faktors, kas ir proporcionāls lokālās polinomu regresijas gludinātāja dispersijai



5.1. att.: Ģenerēti dati ar pievienotu īstās regresijas funkcijas modeli, $\sigma_i = 0.1$, $n = 200$



5.2. att.: Ģenerēti dati ar pievienotu īstās regresijas funkcijas modeli, $\sigma_i = 0.01 + 0.2r(X_i)$, $n = 200$

dizaina punktā x.

4. ticamības joslas pēc San un Loader metodes

$$I_x^{SL} = [\hat{r}(x) - c\hat{\sigma}(x)\|l(x)\|, \hat{r}(x) + c\hat{\sigma}(x)\|l(x)\|].$$

Šeit $\hat{r}(x)$ ir $r(x)$ novērtējums, kas tika iegūts, pielietojot lokālo lineāru gludinātāju ar Gausa kodolu. Gludināšanas parametrs h tika izvēlēts, izmantojot krosvalidācijas procedūru.

Lai novērtētu apskatīto metožu efektivitāti, 1000 reižu tiek simulēti dati pie dažādu izlašu apjomiem. Katrai izlasei tika ģenerētas 1000 butstrapotās izlases, lai noteiktu attiecīgu kvantili. Tiek konstruētas vienlaicīgās ticamības joslas neparametriskajai regresijai pie nozīmības līmeņiem $\alpha = 0.10$, 0.05 un 0.01. Programmā R uzrakstītā algoritma, kas rēķina pārklājuma precizitāti uzkonstruētajām ticamības joslām, darbības laiks ir ļoti liels. Sakarā ar to, ka netika atrasts risinājums, kā var optimizēt doto uzdevumu, tika pieņemts lēmums samazināt butstrapoto izlašu skaitu līdz 350, bet kā gludināšanas parametru h izmantot plug-in novērtējumu, kas ir iebūvēts R programmā kā funkcija *dpill*. Tādā veida tika aprēķinātas pārklājuma precizitātes ticamības joslām. Rezultāti tika apkopoti Tabulās 5.1.,5.2. un 5.3.

Gadījumā, kad dispersija ir konstanta, var novērot, ka pārklājuma precizitātes visām ticamības joslām ir ļoti tuvas pārklājuma teorētiskajām varbūtībām. Tomēr rezultāti ir

5.1. tabula: Pārklājuma precizitāte vienlaicīgajām ticamības joslām pie izlases apjoma $n = 100$.

	Ticamības līmenis	I_x^U	I_x^T	I_x^W	I_x^{SL}
$\sigma_i = 0.1$	90%	0.831	0.850	0.863	0.870
	95%	0.900	0.887	0.902	0.935
	99%	0.940	0.950	0.943	0.975
$\sigma_i = 0.05 + 0.1r(X_i)$	90%	0.845	0.860	0.843	0.610
	95%	0.860	0.892	0.887	0.665
	99%	0.910	0.965	0.939	0.685
$\sigma_i = 0.01 + 0.2r(X_i)$	90%	0.617	0.871	0.826	0.525
	95%	0.757	0.904	0.913	0.560
	99%	0.896	0.957	0.956	0.655

5.2. tabula: Pārklājuma precizitāte vienlaicīgajām ticamības joslām pie izlases apjoma $n = 200$.

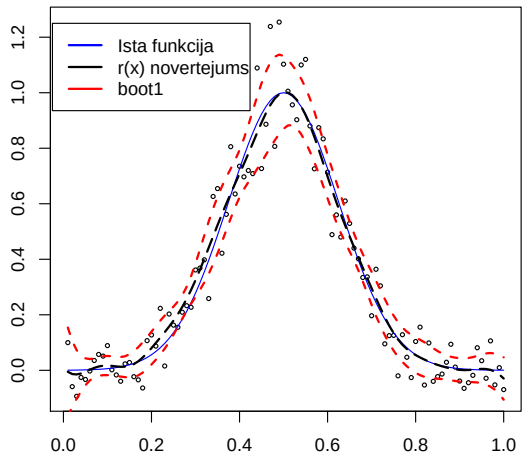
	Ticamības līmenis	I_x^U	I_x^T	I_x^W	I_x^{SL}
$\sigma_i = 0.1$	90%	0.867	0.870	0.857	0.886
	95%	0.900	0.923	0.905	0.927
	99%	0.966	0.971	0.971	0.979
$\sigma_i = 0.05 + 0.1r(X_i)$	90%	0.886	0.895	0.857	0.677
	95%	0.923	0.923	0.914	0.713
	99%	0.942	0.971	0.962	0.795
$\sigma_i = 0.01 + 0.2r(X_i)$	90%	0.762	0.894	0.867	0.559
	95%	0.876	0.905	0.933	0.600
	99%	0.914	0.971	0.981	0.668

nedaudz labāki ticamības joslām I_x^{SL} un I_x^W, kas ir vairāk pielāgotas homoskedastistiskajām gadījumam. Pieaugot heteroskedascitātei, butstrapa ticamības joslām joprojām ir laba pārklājuma precizitāte, kamēr I_x^{SL} joslu pārklājuma kļūda ievērojami pieaug. Starp butstrapa joslām sliktāki rezultāti ir joslām I_x^T, jo šeit tiek ņemta vērā dispersijas novēr-

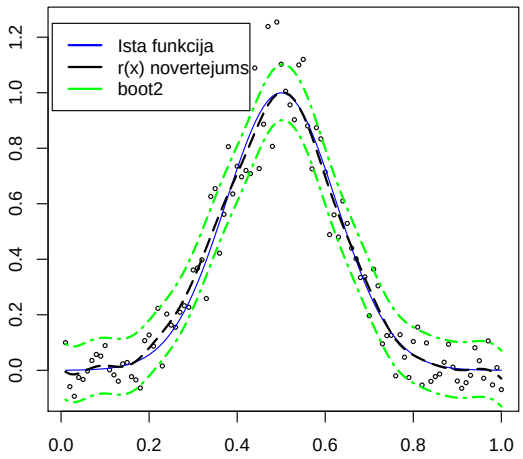
5.3. tabula: Pārklājuma precizitāte vienlaicīgajām ticamības joslām pie izlases apjoma $n = 300$.

	Ticamības līmenis	I_x^U	I_x^T	I_x^W	I_x^{SL}
$\sigma_i = 0.1$	90%	0.898	0.901	0.914	0.893
	95%	0.952	0.962	0.942	0.955
	99%	0.981	0.994	0.991	0.987
$\sigma_i = 0.05 + 0.1r(X_i)$	90%	0.886	0.904	0.901	0.682
	95%	0.953	0.962	0.952	0.732
	99%	0.981	0.980	0.982	0.808
$\sigma_i = 0.01 + 0.2r(X_i)$	90%	0.866	0.914	0.905	0.587
	95%	0.895	0.962	0.952	0.615
	99%	0.962	0.981	0.998	0.696

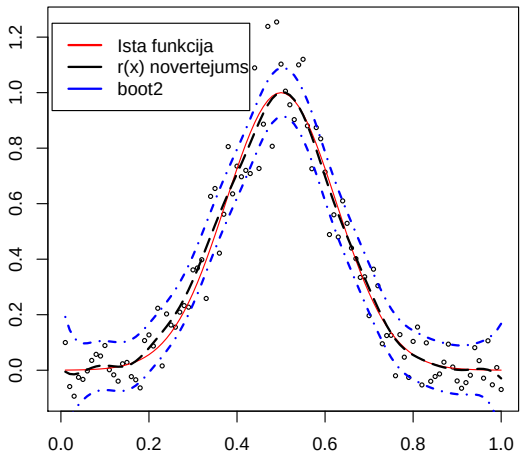
tējuma izkliede, līdz ar to notiek ticamības joslu platuma sašaurinājums. Tajā paša laikā I_x^U un I_x^W joslu platums ir ievērojami lielāks nekā mainīga platuma ticamības joslām. Ilustratīvais piemērs vienai un tai pašai simulācijai tiek atspoguļots attēlos 5.3., 5.4. un 5.5.



5.3. att.: Vienlaicīgās ticamības joslas I_x^T neparametriskajai regresijai, $\sigma_i = 0.05 + 0.1r(X_i)$, $\alpha = 0.05$, $n = 100$



5.4. att.: Vienlaicīgās ticamības joslas I_x^U neparametriskajai regresijai, $\sigma_i = 0.05 + 0.1r(X_i)$, $\alpha = 0.05$, $n = 100$



5.5. att.: Vienlaicīgās ticamības joslas I_x^W neparametriskajai regresijai, $\sigma_i = 0.05 + 0.1r(X_i)$, $\alpha = 0.05$, $n = 100$

6. Pielietojumi

Šajā nodaļā konstruēsim vienlaicīgas ticamības joslas reālai datu problēmai. Tiek apskatīti divu veidu dati: ar homoskedastisku un heteroskedastisku struktūru. Un abos gadījumos tiek pārbaudīta hipotēze par to, ka atlikumu dispersija ir atkarīga no regresora, izmantojot Goldfelda-Kvandta testu.

Goldfelda-Kvandta tests (Goldfeld-Quandt test).

$$H_0 : \ \sigma_i = Const$$

$$H_1 : \ \sigma_i = \sigma(X_i).$$

Testa procedūra ir sekojoša. No sākuma sakārto novērojumu pārus $(X_1, Y_1), \ldots, (X_n, Y_n)$ pēc prediktora vērtībām. Un sadala datus divās grupās (X_I, Y_I) un (X_{II}, Y_{II}) ar attiecīgiem izlases apjomiem n_1 un n_2, kur $n_1 + n_2 = n$. Nulles hipotēze tiek noraidīta, ja atlikumu dispersijas divos lineāros regresijas modeļos ir vienādas. Tātad pieņemsim, ka rezidiju vektori ir $\hat{\epsilon}_1$ un $\hat{\epsilon}_2$, tad testa statistika ir sekojoša

$$R = \frac{\hat{\epsilon}_1^T \hat{\epsilon}_1/(n_1 - 1)}{\hat{\epsilon}_2^T \hat{\epsilon}_2/(n_2 - 1)}.$$

$R \xrightarrow{d} F_{(n_1-p),(n_2-p)}$, ja H_0 ir spēkā. Šī testa procedūra tiek domāta dilstošai dispersijai. Augošai dispersijai jāmaina vietām indeksi testa statistikā. Eksistē šī testa modifikācija. Jāizslēdz aptuveni $n/4 = d$ (dažos literatūras avotos piedāvā izslēgt $n/3$ novērojumus) vidējās vērtības no sakārtotās datu kopas. Pēc tam jāveic regresija pēc pirmajiem $n/2-d/2$ novērojumiem un pēc pēdējiem $n/2 - d/2$ novērojumiem, iegūstot atlikumus $\hat{\epsilon}_1$ un $\hat{\epsilon}_2$. Testa statistika ir sekojoša

$$R = \frac{\hat{\epsilon}_1^T \hat{\epsilon}_1}{\hat{\epsilon}_2^T \hat{\epsilon}_2}.$$

$R > F_{(\frac{n}{2}-\frac{d}{2}-2),(\frac{n}{2}-\frac{d}{2}-2)}$, ja H_0 ir spēkā.

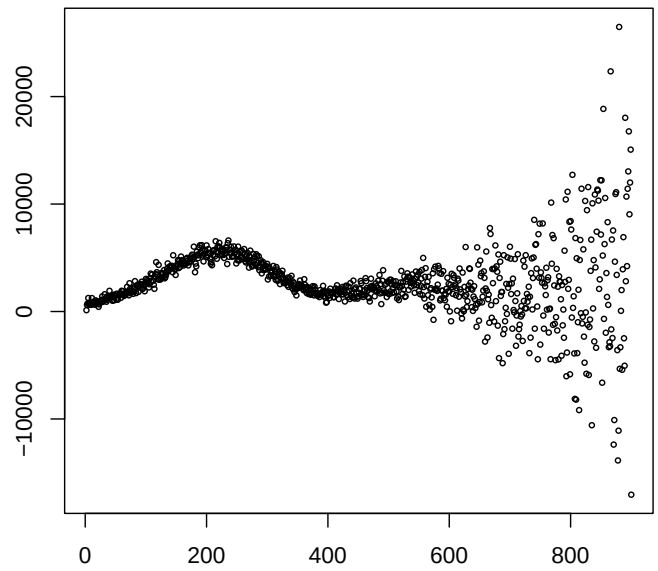
Ja mēs izslēgsim pārāk daudz novērojumu, tad kvadrātu atlikumu summām RSS_1 un RSS_2 būs pārāk zemas brīvības pakāpes. Ja mēs izslēgsim pārāk maz novērojumu, tad testa efektivitāte būs ļoti zema, jo RSS_1 un RSS_2 salīdzināšana būs neefektīva. Sīkāk par šo testu skatīt [5].

Mēs nelietosim šī testa modifikāciju, jo īsti nav skaidrs kā izvēlēties izslēgto novērojumu skaitu. Goldfelda-Kvandta tests tiek iebūvēts programmā R kā funkcija *gqtest*.

Tātad no sākuma apskatīsim heteroskedastisko datu piemēru, kas pilnā mērā ilustrēs šī darba rezultātus.

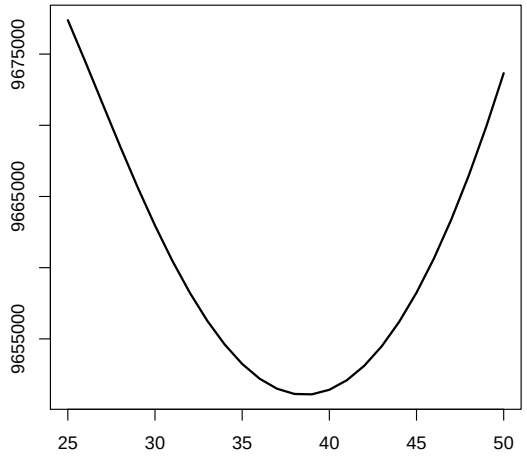
Heteroskedastisko datu piemērs

Tiek analizēti CMB(cosmic microwave background radiation) dati - novērotais fluktuāciju spēks atkarībā no temperatūras fluktuācijas frekvences no Larry Wassermana grāmatas [3]. Datus var brīvi lejuplādēt no mājaslapas http://www.stat.cmu.edu/larry/all-of-nonpar/data.htm. Novērojumu skaits ir pietiekoši liels, $n = 899$. CMB dati tika atspoguļoti Attēlā 6.1.

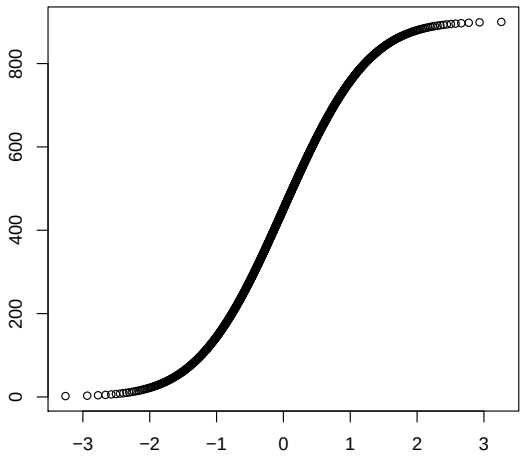


6.1. att. Fluktuāciju spēks atkarībā no temperatūras fluktuācijas frekvences

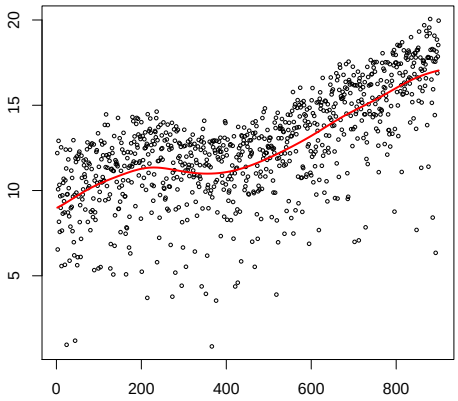
Pēc CMB datu grafika var redzēt, ka dispersija kļūst lielāka, pieaugot mainīgā X_i



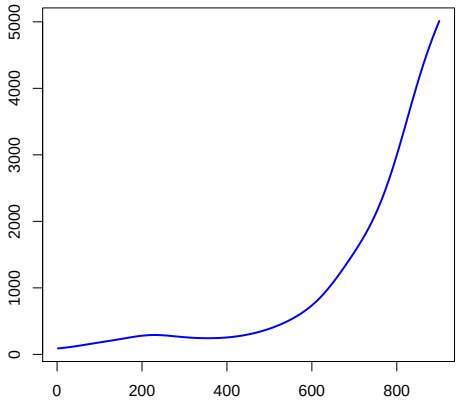
6.2. att. $\hat{R}(h)$ atkarībā no joslas platuma



6.3. att. Q-Q grafiks



6.4. att. Lokālā regresija Z_i pret X_i



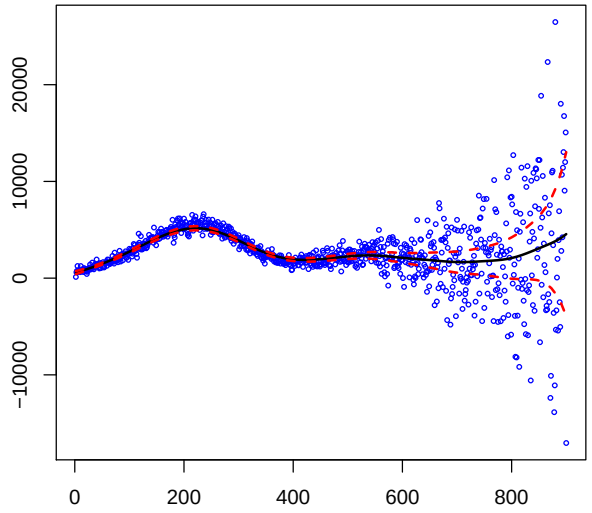
6.5. att. Dispersijas novērtējums

vērtībai. Veicot Goldfelda-Kvandta testu pie nozīmības līmeņa $\alpha = 0.05$, tika noraidīta nulles hipotēze par analizējamo datu homogenitāti, tā kā $p-$vērtība ir mazāka par $2.2e-16$.

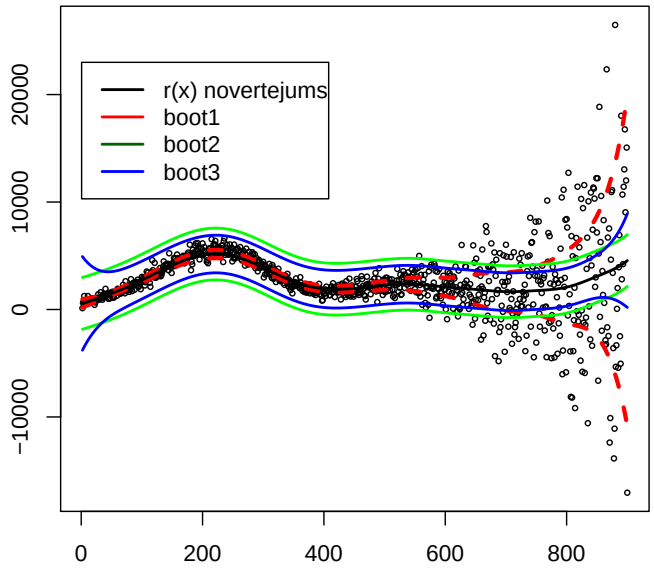
Lai novērtētu regresijas funkciju, sākumā jāizvēlas gludināšanas parametru h. Jāminimizē krosvalidācijas funkcija (2.3.1). Attēlā 6.2. tika atspoguļots šīs funkcijas grafiks.

Tātad ar krosvalidācijas palīdzību iegūtais optimālais joslas platums ir $h = 39$, kas kontrolē gludināšanas pakāpi.

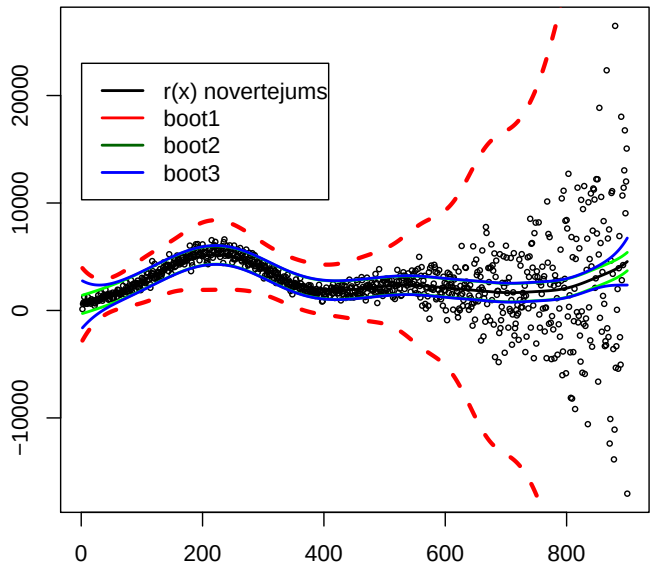
Izmantojot doto optimālo joslas platumu un pielietojot lokālo lineāru gludinātāju ar



6.6. att.: Vienlaicīgas ticamības joslas I_x^{SL} neparametriskajai regresijai, $\alpha = 0.05$, CMB dati



6.7. att.: Vienlaicīgas ticamības joslas neparametriskajai regresijai, izmantojot mežonīgo butstrapu, $\alpha = 0.05$, CMB dati



6.8. att.: Vienlaicīgas ticamības joslas neparametriskajai regresijai, izmantojot neparametrisko rezidiju butstrapu, $\alpha = 0.05$, CMB dati

Gausa kodolu, tika iegūts regresijas novērtējums $\hat{r}(x)$.

Tā kā San un Loadera modelim ir pieņēmums par atlikumu normalitāti, jāpārbauda vai tie ir normāli sadalīti.

Attēlā 6.3. redzamais Q-Q grafiks liecina par to, ka atlikumiem nav normālais sadalījums. Arī veicot datu normalitātes pārbaudi ar Kolmogorova-Smirnova testa palīdzību, jāsecina, ka, tā kā $p-$vērtība ir mazāka par $2.2e-16$ tad, pie nozīmības līmeņa $\alpha = 0.05$ jānoraida hipotēze par analizējamo datu normalitāti.

Nākamais solis ir dispersijas $\sigma(X_i)$ novērtēšana. Tāpēc tiek aprēķinātas vērtības $Z_i = \ln(Y_i - \hat{r}(X_i))^2$ un veikta lokāla regresija Z_i pret X_i. Šīs regresijas funkcijas grafiks tiek atspoguļots Attēlā 6.4.

Ar šīs regresijas palīdzību tiek iegūts novērtējums $\hat{q}(x)$. Savukārt, izmantojot iepriekšējo novērtējumu, tiek aprēķināts dispersijas novērtējums $\hat{\sigma}^2(x) = \exp(\hat{q}(x))$. Attēlā 6.5. var redzēt dispersijas novērtējumu atkarībā no regresora.

Dispersija tiešām nav konstants lielums, jo, pieaugot mainīgā X vērtībai, variācija kļūst tikai lielāka.

Tādējādi uz doto brīdi ir visa nepieciešamā informācija, lai uzkonstruētu vienlaicīgas

ticamības joslas I_x^{SL}, I_x^T, I_x^U un I_x^W, kas tika definētas 6.nodaļā. Sākumā aprēķināsim konstantes k_0 un c, kas tiek izmantotas San un Loader joslu konstruēšanai. Pēc formulas (4.1.4) tiek aprēķināta konstante $k_0 = 36.11$. Ar iegūtās konstantes palīdzību pēc formulas (4.1.5) tiek aprēķināts $c = 3.30571$. Uzkonstruētās vienlaicīgas ticamības joslas I_x^{SL} pie nozīmības līmeņa $\alpha = 0.05$ tiek atspoguļotās Attēlā 6.6.. Savukārt joslas I_x^T, I_x^U un I_x^W (uz grafika attiecīgi *boot*1, *boot*2, *boot*3) tiek konstruētas, izmantojot 1000 bustrapa izlases, Attēls 6.7.

Analizējot uzkonstruēto vienlaicīgo ticamības joslu grafikus, var secināt, ka San un Loader un *boot*1 joslas ir jūtīgas pret datu izkliedi, kamēr *boot*2 un *boot*3 joslas ir plašākas un ar vienmērīgu joslas platumu.

Uzkonstruēsim ticamības joslas *boot*1, *boot*2 un *boot*3, pielietojot rezidiju butstrapu (Attēls 6.8.)

Analizējot ar rezidiju butstrapu uzkonstruētās joslas, var secināt, ka tiešām rezidiju butstraps slikti strādā uz heteroskedastiskiem datiem. *Boot*1 un *boot*2 joslas ir pārāk šauras. Savukārt *boot*1 kļūst pārāk jūtīgs pret datu izlecējiem.

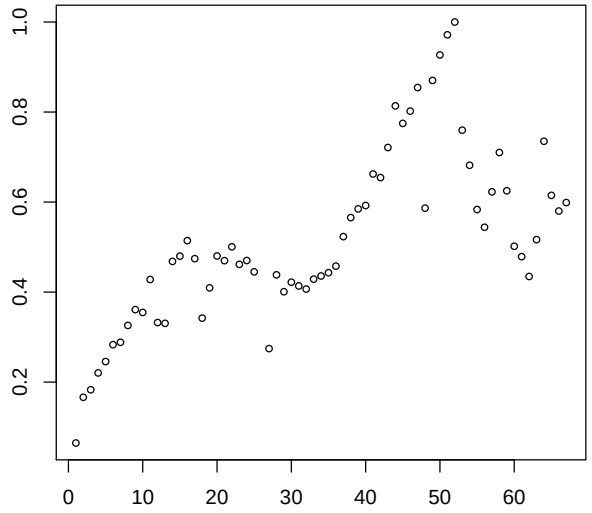
Homoskedastisko datu piemērs

Dati ar homoskedastisku struktūru tika iegūti no apdrošināšanas kompānijas "ERGO Latvija Dzīvība". Tā kā datu izmantošana ārpus kompānijas nav atļauta, sākotnējiem datiem tika mainīta skala.

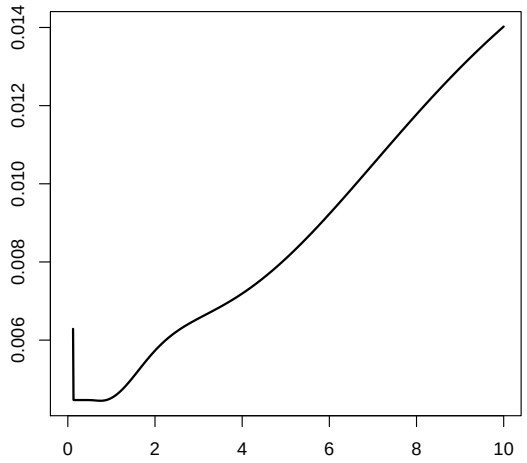
Apskatīsim grupas klienta veselības apdrošināšanā pieteikto atlīdzību skaitu atkarībā no polises lietošanas ilguma. Parasti šāda tipa dati tiek izmantoti apdrošināšanas produkta tarifu noteikšanā. Laiks tiek mērīts nedēļās. Novērojumu skaits ir neliels, $n = 66$. Dati tika atspoguļoti Attēlā 6.9. Veicot Goldfelda-Kvanda testu, nulles hipotēze par analizējamo datu homogenitāti netiek noraidīta pie nozīmības līmeņa $\alpha = 0.05$, tā kā $p-$vērtība ir vienāda ar 0.6884.

Tālāk ar krosvalidācijas metodes palīdzību izvēlēsimies gludināšanas parametru h. Iegūtais optimālais joslas platums ir $h = 0.75$. Izmantojot doto optimālo joslas platumu un pielietojot lokālu lineāru gludinātāju, tika iegūts regresijas novērtējums $\hat{r}(x)$. Nākamais solis ir konstantas dispersijas novērtēšana pēc Teorēmas 6. Iegūstam, ka $\hat{\sigma}^2 = 0.06289096$. Tālāk tiek pārbaudīta atlikumu normalitāte.

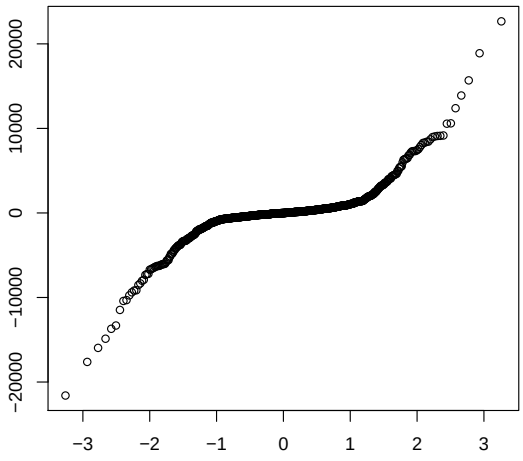
Pēc Attēla 6.11. var iedomāties, ka atlikumi tiek sadalīti normāli. Veicot Kolmogorova-



6.9. att. Pieteikto atlīdzību skaits atkarībā no polises lietošanas ilguma



6.10. att. $\hat{R}(h)$ atkarība no joslas platuma

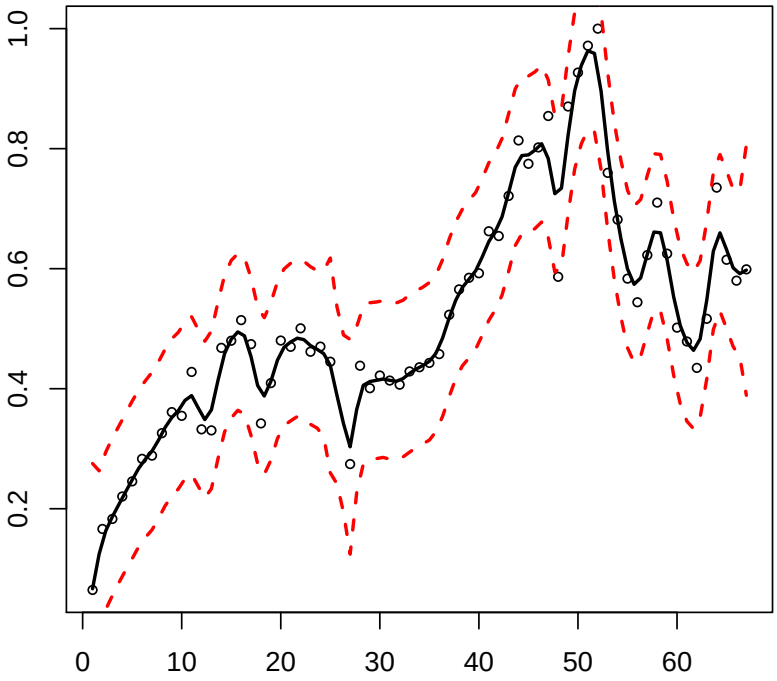


6.11. att. Q-Q grafiks

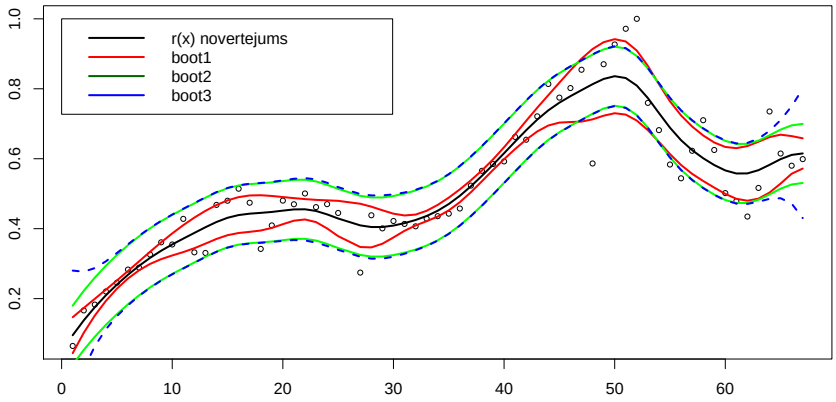
Smirnova testu, jāsecina, ka, tā kā $p-$vērtība ir vienāda ar 0.1099, tad pie nozīmības līmeņa $\alpha = 0.05$, hipotēze par analizējamo datu normalitāti netiek noraidīta.

Aprēķinot konstantes k_0 ($k_0 = 48.63$) un c ($c = 3.39$), ar San un Loader metodes palīdzību tika konstruētas vienlaicīgās ticamības joslas (Attēls 6.12.). Un, izmantojot 1000 mežonīgā butstrapa izlases, tika iegūtas ticamības joslas I_x^T, I_x^U, I_x^W (Attēls 6.13.).

Analizējot uzkonstruēto vienlaicīgo ticamības joslu grafikus, var secināt, ka I^{SL}, *boot*2

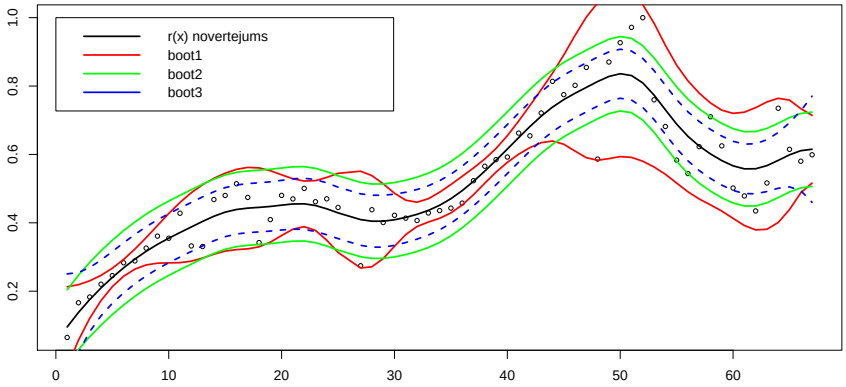


6.12. att.: Vienlaicīgas ticamības joslas I_x^{SL} neparametriskajai regresijai, $\alpha = 0.05$, ERGO kompānijas dati



6.13. att.: Vienlaicīgas ticamības joslas neparametriskajai regresijai, izmantojot mežonīgo butstrapu, $\alpha = 0.05$, ERGO kompānijas dati

un *boot*3 joslas pēc platuma un uzvedības ir ļoti līdzīgas. Ticamības joslas *boot*1 ir ļoti jūtīgas pret izlecējiem, tāpēc var secināt, ka šī metode labāk strādā pie hereskedastiskās struktūras. Salīdzinājumam uzkonstruēsim vienlaicīgās ticamības joslas *boot*1, *boot*2 un *boot*3, izmantojot rezidiju butstrapu. Analizējot ar rezidiju butstrapu uzkonstruētās jos-



6.14. att.: Vienlaicīgas ticamības joslas neparametriskajai regresijai, izmantojot neparametrisko rezidiju butstrapu, $\alpha = 0.05$, ERGO kompānijas dati

las, var secināt, ka rezidiju butstraps dod diezgan plašas ticamības joslas *boot*2 un *boot*2 gadījumos. Un kārtējo reizi var pārliecināties, ka *boot*1 jāpiemēro heteroskedastiskiem datiem, it īpaši, ja izlases apjoms nav liels, kā arī ir šajā gadījumā. Homoskedastiskiem datiem joslas *boot*1 praktiski nav pielietojamas.

Secinājumi

Vienlaicīgo ticamības joslu konstruēšanai neparametriskajā regresijā tika apskatītas divas pieejas: klasiskā San un Loader un rezidiju butstrapa metodes. Sākotnējais uzdevums bija veikt San un Loader metodes detalizētāku analīzi. Pats grūtākais posms vienlaicīgo ticamības joslu konstruēšanā neparametriskajai regresijai ir konstantes k_0 aprēķins. No k_0 ir atkarīgs konstantes c aprēķins, kas ietekmē ticamības joslu konstruēšanu. Konstante k_0 tika uzprogrammēta programpaketē R.

Literatūrā ir zināms, ka eksistē rezidiju butstrapa dažādi veidi. Darbā tika veikta butstrapa metožu analīze un salīdzinājums. No tā tika secināts, ka mežonīgais butstraps vislabāk ir piemērots heteroskedastiskiem datiem. Un turpmāk regresijas modeļos tika apskatīts rezidiju mežonīgais butstraps. Nākamais solis bija ar rezidiju mežonīgā butstrapa palīdzību uzkonstruēt vienlaicīgās ticamības joslas neparametriskajai regresijai.

Lai apstiprinātu iegūtos teorētiskos rezultātus, tika salīdzinātas vienlaicīgās ticamības joslas ar simulāciju metodes palīdzību. Simulējot datus pie dažādas dispersijas struktūras, tika uzkonstruētas vienlaicīgas ticamības joslas un, izmantojot pārklājuma precizitāti, tika analizēta metožu efektivitāte. Simulācijās iegūtie rezultāti liecina, ka heteroskedastisko datu gadījumā, klasiskā pieeja dod šaurākas joslas nekā mežonīga butstrapa metode. Savukārt homoskedastisko datu gadījumā, San un Loader metode dod nedaudz labākus rezultātus. Palielinoties izlases apjomam, visām četrām ticamības joslām pārklājuma precizitātes konverģē uz teorētiskajām, tomēr samērā lēni.

Viens no darba uzdevumiem bija pielietot apskatītās metodes reālai problemātikai. Tika apskatīti heteroskedastiskie un homoskedastiskie dati. Heteroskedastisko datu gadījumā tika izmantoti CMB (cosmic microwave background radiation) dati, iegūtie no Larry Wassermanna mājas lapas. Ar mežonīgā butstrapa metodi konstruētas ticamības joslas neparametriskajai regresijai pie nozīmības līmeņa $\alpha = 0.05$ ir plašākas un labāk pielāgotas datu struktūrai nekā joslas, kas tika iegūtas ar San un Loader palīdzību. Konstruējot tās pašās ticamības joslas, bet tikai pielietojot neparametrisko rezidiju butstrapu nevis mežonīgu butstrapu, tika secināts, ka tiešām neparametriskais rezidiju butstraps dod šaurākas ticamības joslas, kas mazāk ievēro datu heteroskedastisku dabu.

Homoskedastisko datu gadījumā tika izmantoti dati no apdrošināšanas kompānija "Ergo Latvija Dzīvība" par viena grupas klienta, kam ir veselības apdrošināšana, pieteikto atlīdzību skaitu atkarībā no apdrošināšanas polises lietošanas ilguma. Šajā gadījumā ar

San un Loader metodes palīdzību uzkonstruētās joslas ir ļoti plašas, tam viens no iemesliem ir neliels izlases apjoms. Ar mežonīgo butstrapu uzkonstruētas ticamības joslas ir šaurākas. Savukārt ar neparametrisko bustrapu uzkonstruētas joslas ir plašākas nekā ar mežonīgo bustrapu uzkonstruētās joslas.

Izmantotā literatūra un avoti

[1] J. Sun and C. Loader. Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.*, 22:1328–1345, 1994.

[2] C. F. J. Wu. Jacknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14:1261–1295, 1986.

[3] Larry Wasserman. *All of nonparametric statistics.* Springer, 2003.

[4] http://mars.wiwi.hu-berlin.de/ebooks/html/spm/.

[5] Jurgen Groβ. *Linear regression.* Springer, 2003.

[6] Bradley Efron and Robert J. Fibshirani. *An introduction to the bootstrap.* CHAPMAN HALL/CRC, 1993.

[7] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9:1196–1217, 1981.

[8] W. Hardle and E. Mammen. Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21:1926–1947, 1993.

[9] E. Flachaire. Bootstrapping heteroskedastic regression models: Wild bootstrap vs pairs bootstrap. *Computational Statistics and Data Analysis, forthcoming*, 2004.

[10] R. Davidson and E. Flachaire. The wild bootstrap, tamed at last. *GREQAM Document de Travail 99A32, revised*, 2001.

[11] J. G. MacKinnon. Bootstrap inference in econometrics. *Canad. J. Econom.*, 35:615–645, 2002.

[12] D. Q. Naiman. Volumes of tubular neighborhoods of spherical polyhedra and statistical inference. *Ann. Statist.*, 18:685–716, 1990.

[13] J. Faraway and J. Sun. Simultaneous confidence bands for linear regression with heteroscedastic errors. *J. Amer. Statist. Assoc.*, 90:1094–1098, 1995.

[14] Raz J. Sun J. and Faraway J. J. Confidence bands for growth and response curves. *Statistica Sinica*, 3:679–698, 1999.

[15] R. L. Eubank and P. L. Speckman. Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.*, 88:1287–1301, 1993.

[16] P. Bickel and M. Rosenblatt. On some global measures of the deviations of density function estimates. *Ann. Statist.*, 1:1071–1096, 1973.

[17] M. H. Neumann and J. Polzehl. Simultaneous bootstrap confidence bands in nonparametric regression. *Nonparametric Statistics*, 9:307–333, 1998.

A Pielikums

A1. Programmas kods vienlaicīgo ticamības joslu konstruēšanai neparametriskajai regresijai ar San un Loader metodes palīdzību

```
visi.dati<-read.table(file="cmb.txt", header=T)
x.data<-visi.dati[1:length(visi.dati[,1]),1]
y.data<-visi.dati[1:length(visi.dati[,1]),2]
n<-length(x.data)
y.data
x.data
library(KernSmooth)
h<-38.57
#Lokala regresija( rn(x).n=sum(li(x)*Yi); lj(x)=bj(x)/sum bj; bi(x)=K(xi-x/h)*(Sn,2(x)-(xi-x)Sn,1(x))
## Sn,j=sum(K(xi-x/h)*(xi-x)^j )
###                  Gausa kodols
kod<-function(x){
1/sqrt(2*pi)*exp(-x^2/2)
}
###                Funkcija Snj

Snj<-function(x,j,dati,h){
sum(kod((dati-x)/h)*(dati-x)^j)
}
###                   Funkcija bi(x)
bi<-function(x,i,dati,h){
kod((dati[i]-x)/h)*(Snj(x,2,dati,h)-(dati[i]-x)*Snj(x,1,dati,h))
}
###                   Funkcija li(x)
li<-function(x,i,dati,h){
n<-length(dati)
bi(x,i,dati,h)/sum(bi(x,1:n,dati,h))
}
###                    Funkcija Ti(x)
Ti<-function(x,i,dati,h){
kk<-sum(bi(x,1:n,dati,h))
```

```
l.norma<-sqrt(sum(li(x,1:n,dati,h)^2))
li(x,i,dati,h)/l.norma
}
###                      Zi punkti
k<-10000
min.x<-min(x.data)
max.x<-max(x.data)
m<-(max.x-min.x)/k
z<-seq(min.x,max.x,by=m)
mm<-length(z)
mm
                       Matrica T
T.matr<-matrix(0,length(z),n)
for (i in 1:length(z))
{
 T.matr[i,1:n]<-Ti(z[i],1:n,x.data,h)
}
###                   Matrica T(i)-T(i-1)
matr.pedeja<-T.matr[2:mm,1:n]-T.matr[1:(mm-1),1:n]
matr.st<-matrix(0,length(z),n)
for (i in 2:length(z))
{
 matr.st[i,1:n]<-(T.matr[i,1:n]-T.matr[i-1,1:n])^2
}
normas<-matrix(0,1,n)
for(i in 1:n)
{
normas[1:1,i]<-sqrt(sum(matr.st[1:length(z),i]))
}
k0<-sum(normas[1,1:n])
###                  Gausa kodola atvasinajums
kod_atv<-function(x){
(-1/sqrt(2*pi))*x*exp(-x^2/2)
}
###                 c0 aprekins
alfa<-0.05
k_n<-k0
g<-function(x){
alfa-2*(1-pnorm(x,0,1))-(k_n/pi)*exp(-x^2/2)
}
cc<-uniroot(g,c(-100,100))
c_const<-cc$root
###              Lokalas regresijas novertejums rn(x)
rn.loc<-function(x,xdat,ydat,h){
rn.vec<-c()
n<-length(xdat)
      for(j in 1:length(x)){
rn.vec[j]<-sum(li(x[j],1:n,xdat,h)*ydat)
}
```

```
rn.vec
}
###                          Dispersijas novērtējums kā funkcijas
Zi<-log((y.data-rn.loc(x.data,x.data,y.data,h))^2)
xx<-seq(min(x.data),max(x.data),len=100)
h.rez<-dpill(x.data,Zi)
plot(x.data,Zi,xlab="apgabals",ylab="Zi", main="Regresija Zi pret Xi",cex=0.5)
points(xx,rn.loc(xx,x.data,Zi,h.rez),type="l",lwd=2,col="red")

#### Vienlaicīgas ticamības joslas (dispersija NAV konstanta)
bands.loc<-function(x,xdat,ydat,h,cc,zdat){
n<-length(xdat)
l.norm<-sqrt(sum(li(x,1:n,xdat,h)^2))
h.rez<-dpill(xdat,zdat)
sigma<-sqrt(exp(rn.loc(x,xdat,zdat,h.rez)))
saknite<-sqrt(sum((li(x,1:n,xdat,h)*sigma)^2))
c(rn.loc(x,xdat,ydat,h)-cc*sigma*l.norm,rn.loc(x,xdat,ydat,h)+cc*sigma*l.norm)
}
bands<-matrix(rep(0,1*n),2,n)   #salikam nulles iek?? matric?, ir 2 rindas un nn kolonnas liela matrica ar null?m
for(j in 1:n){
bands[1:2,j]<-bands.loc(x.data[j],x.data,y.data,h,c_const,Zi)
}
```

A2. Programmas kods vienlaicīgo ticamības joslu konstruēšanai neparametriskajai regresijai ar mežonīgā butstrapa metodes palīdzību

```
visi.dati<-read.table(file="cmb.txt", header=T)
x.data<-visi.dati[1:length(visi.dati[,1]),1]
y.data<-visi.dati[1:length(visi.dati[,1]),2]
n<-length(x.data)
library(KernSmooth)
##                 Gausa kodols
kod<-function(x){
1/sqrt(2*pi)*exp(-x^2/2)
}
###                      Lokala regresija
###                      Funkcija Snj
Snj<-function(x,j,dati,h){
sum(kod((dati-x)/h)*(dati-x)^j)
}
###                   Funkcija bi(x)
bi<-function(x,i,dati,h){
kod((dati[i]-x)/h)*(Snj(x,2,dati,h)-(dati[i]-x)*Snj(x,1,dati,h))
```

```
}
###                  Funkcija wi(x)
wi<-function(x,i,dati,h){
n<-length(dati)
bi(x,i,dati,h)/sum(bi(x,1:n,dati,h))
}
###             Lokalas regresijas novertejums rn(x)
m_n<-function(x,xdat,ydat,h){
 rn.vec<-c()
 n<-length(xdat)
      for(j in 1:length(x)){
rn.vec[j]<-sum(wi(x[j],1:n,xdat,h)*ydat)
}
 rn.vec
}
###             Novert. epsiloni= Yi-m_nov
en<-y.data-m_n(x.data,x.data,y.data,h)
###   sqrt( mi^(x))
mi.n<-function(x){
 mi.vec<-c()
      for(j in 1:length(x)){
mi.vec[j]<-sqrt(sum((wi(x[j],1:n,x.data,h)^2)*(en^2)))
}
 mi.vec
}
alpha<-0.05
##################  Bootstrapping_parametric ~rbinom
B1<-1000
m_n.b1<-matrix(0,n,B1,byrow=FALSE)
for (i in 1:B1){
      bin<-sample(rbinom(n,1,0.5),replace=FALSE)
       zimes<-c()
      for(k in 1:n){
         if(bin[k]<=0){zimes[k]<--1}
              else {zimes[k]<-1}
               }
residuals<-en*zimes
for(j in 1:n){
          m_n.b1[j,i]<-abs(sum(wi(x.data[j],1:n,x.data,h)*residuals))
}
}
### U*
U.b1<-c()
for(i in 1:B1){
U.b1[i]<-max(m_n.b1[1:n,i]/mi.n(x.data))
}
t.b1<-quantile(U.b1,1-alpha)
##Confidence bands (1.example)
bands.b1<-function(x){
```

```
c(m_n(x,x.data,y.data,h)-t.b1*mi.n(x),m_n(x,x.data,y.data,h)+t.b1*mi.n(x))
}
bands1<-matrix(rep(0,1*n),2,n)
for(j in 1:n){
bands1[1:2,j]<-bands.b1(x.data[j])
}
m_n.b2<-matrix(0,n,B1,byrow=FALSE)
for (i in 1:B1){
      bin<-sample(rbinom(n,1,0.5),replace=FALSE)
       zimes<-c()
      for(k in 1:n){
         if(bin[k]<=0){zimes[k]<--1}
              else {zimes[k]<-1}
               }
residuals<-en*zimes
 for(j in 1:n){
          m_n.b2[j,i]<-abs(sum(wi(x.data[j],1:n,x.data,h)*residuals))
}
}
### U*
U.b2<-c()
for(i in 1:B1){
U.b2[i]<-max(abs(m_n.b2[1:n,i]))
}
t.b2<-quantile(U.b2,1-alpha)
#Confidence bands (2.example)
bands.b2<-function(x){
c(m_n(x,x.data,y.data,h)-t.b2,m_n(x,x.data,y.data,h)+t.b2)
}
bands2<-matrix(rep(0,1*n),2,n)
for(j in 1:n){
bands2[1:2,j]<-bands.b2(x.data[j])
}
###   sqrt( w^(x))
w<-function(x){
 w.vec<-c()
      for(j in 1:length(x)){
w.vec[j]<-sqrt(sum((wi(x[j],1:n,x.data,h)^2)))
}
 w.vec
}
m_n.b3<-matrix(0,n,B1,byrow=FALSE)
for (i in 1:B1){
      bin<-sample(rbinom(n,1,0.5),replace=FALSE)
       zimes<-c()
      for(k in 1:n){
         if(bin[k]<=0){zimes[k]<--1}
              else {zimes[k]<-1}
               }
```

```
residuals<-en*zimes
 for(j in 1:n){
          m_n.b3[j,i]<-abs(sum(wi(x.data[j],1:n,x.data,h)*residuals))
}
}
### U*
U.b3<-c()
for(i in 1:B1){
U.b3[i]<-max(m_n.b3[1:n,i]/w(x.data))
}
t.b3<-quantile(U.b3,1-alpha)
##Confidence bands (3.example)

bands.b3<-function(x){
c(m_n(x,x.data,y.data,h)-t.b3*w(x),m_n(x,x.data,y.data,h)+t.b3*w(x))
}
bands3<-matrix(rep(0,1*n),2,n)
for(j in 1:n){
bands3[1:2,j]<-bands.b3(x.data[j])
}
plot(x.data,y.data,col="black",xlab=" ", ylab="",cex=0.5)
points(x.data,bands1[2,],type="l",col="red",lwd=3,lty=2)
points(x.data,bands1[1,],type="l",col="red",lwd=3,lty=2)
points(x.data,m_n(x.data,x.data,y.data,h),type="l",lwd=2, lty=1)
points(x.data,bands2[2,],type="l",col="green",lwd=2,lty=1)
points(x.data,bands2[1,],type="l",col="green",lwd=2,lty=1)
points(x.data,bands3[2,],type="l",col="blue",lwd=2,lty=1)
points(x.data,bands3[1,],type="l",col="blue",lwd=2,lty=1)
legend(0.65,23000,c("r(x) novertejums","boot1","boot2","boot3"),lwd=2,col=c("black","red","darkgreen","blue"))
```

A3. Programmas kods joslas platuma aprēkinam ar krosvalidācijas palīdzību

```
visi.dati<-read.table(file="cmb.txt", header=T)
x.data<-visi.dati[1:length(visi.dati[,1]),1]
y.data<-visi.dati[1:length(visi.dati[,1]),2]
n<-length(x.data)
#####         Nadaraya-Watson kodolu novērtējums
nad.fun<- function(x){sum(dnorm((x-x.data)/h)*y.data)/sum(dnorm((x-x.data)/h))}
###                            Krosvalidācija
l.nad<-function(x,i,h){dnorm((x-x.data[i])/h)/sum(dnorm((x-x.data)/h))}
rez<-c()
hh<-seq(0,50,by=0.01)
for (j in 1:length(hh)){
h<-hh[j]
ss<-0
for (i in 1:n){
ss<-ss+((y.data[i]-nad.fun(x.data[i]))/(1-l.nad(x.data[i],i,h)))^2
```

```
}
rez[j]<-ss/n
}
plot(hh,rez,type="l",lwd=2,xlab="",ylab="" )
index<-order(rez)[1]
order(rez)
h.opt<-hh[index] ### minimizējošais h
h.opt
```

Maģistra "Vienlaicīgās ticamības joslas neparametriskajā regresijā" 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: Nataļja Saveļjeva

_____ _____
(paraksts) (datums)

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

_____ _____
(paraksts) (datums)

Recenzents: Asociētā profesore Viktorija Carkova

_____ _____
(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____
(datums)

_____
(darbu pieņēma)

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____
(datums)

Komisijas sekretārs/-e: _____ _____
(Vārds, Uzvārds) (paraksts)