LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

LOKĀLĀ TICAMĪBAS FUNKCIJA

Kursa darbs

Autors: **Oļesja Grigorčaka**
Stud. apl. og05009
Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2009

Anotācija

Kursa darbā tika aplūkota lokālās ticamības funkcija un tās pielietojums blīvuma funkcijas novērtējumam un regresijas modelim. Lokālā aproksimācija ar ticamības funkciju tiek veikta pie dažādām polinoma pakāpēm $p \geq 0$. Maksimizējot vislielākās ticamības funkciju pie $p \geq 1$ atrisinājumam nav analītiskas formas. Tiek aplūkoti blīvuma funkcijas novērtējuma nepieciešamie nosacījumi un eksistence. Tiek veikta Monte-Karlo simulācijas un prakstiska datu analīze ar *Locfit* paketi programmā R. Darbā ir iesākts aplūkot lokālās ticamības funkcijas regresijas modeli, tā praktiskā datu analīze, ko izvērstāk varētu analizēt, rakstot diplomdarbu.

Atslēgas vārdi: lokālā ticamības funkcija; blīvuma funkcijas novērtējums; lokālā regresija

Saturs

Ievads

Pirmoreiz Brillinger (1977) veica lokālās ticamības funkcijas aproksimācijas pielietojumu. Vēlāk studēja arī Tishbirani (1984), viņa darbā tika apskatīts lokālās ticamības funkcija un tās pielietojums Cox, Logistic modelim.[1] Lokālā regresija tika apskatīta speciālgadījumā ar lokālas ticamības funkciju ir aprakstīta Tibshirani un Hastie (1987). Apskatītais ir paredzēts neparametriskās regresijas modelēšanai situācijās, piemēram, loģistikas regresijas un proporcionālajiem Hazard modeļiem[2]. Staniswalis (1989) var skatīt par lokālās ticamības funkcijas pielietojumu Cox modelim.[3]

Lokālā polinoma metodes sadalījuma novērtējums ir aprakstīts autoru Lejeune un Sarda (1992) darbā. Šajā metodē sadalījuma formulēšanai sadalījuma funkcija tiek novērtēta, izmantojot svērto kvadrātisko formu un tiek izmantots logaritms no blīvuma funkcijas.[4]

Hjort un Jones (1996) ir pētījuši lokālās ticamības funkcijas pielietojumu blīvuma funkcijas novērtējumam un autori balstās, galvenokārt, uz viendimensijas gadījumu, taču neizmanto lokālo log-polinoma formu. Autori apraksta kā atrast būtiskas īpašības,izmantojot lokālā modeļa parametru skaitu nevis precīzu modeļa formu. Hjort un Jones apraksta lokālās ticamības funkcijas pielietojumu ar citām metodēm. Hjort (1997) apraksta lokālas ticamības funkcijas pielietojumu Hazard modelim.

Lokālās ticamības metodes risina lokālās optimizācijas problēmas, izmantojot novirzīto dispersiju.

Kursa darba 1.daļā tiek apskatīta log-ticamības funkcija, lokālā log-ticamības funkcija un tās pielietojums blīvuma funkcijas novērtēšanā. Tiek apskatīta blīvuma funkcijas novērtējuma eksistences un unitātes nosacījumi. Aproksimācijas rezultāts bieži vien ir atkarīgs no kodola h izvēles joslas platuma izvēles. Programmā *Locfit* esmu parādījusi blīvuma funkcijas novērtēšanu atkarībā no h izvēles. Ir apskatīts blīvuma funkcijas novērtējums pie $p = 0$ un $p = 1$ polinoma pakāpēm, kur pie $p \geq 1$ atrisinājumam nav analītiskas formas.

Kursa darba 2.daļā tika pētīts lokālās regresijas modeļa log-ticamības funkcija, lokālā logistic regresija izmantojot Bernulli regresijas modeli. Tiek izklāstīts kapēc mums vajag maksimizēt lokālo ticamības funkciju.

Tiek aplūkota krosvalidācijas metode.

1. Blīvuma funkcijas novērtējums

1.1. Lokālā ticamības funkcija blīvuma novērtējumam

Pieņem, ka gadījuma lielumi $X_1, X_2..., X_n$ ir vienādi sadalīti neatkarīgi gadījuma lielumi ar blīvuma funkciju f.

Aplūko log-ticamības funkciju

$$\mathcal{L}(f) = \sum_{i=1}^{n} log(f(X_i)) - n(\int_{\mathcal{X}} f(u)du - 1), \tag{1.1.1}$$

kur $\mathcal{X}$ ir blīvuma funkcijas intervāla robežas. Saskaitāmais $n(\int_{\mathcal{X}} f(u)du - 1)$ parasti neietilpst lokālās ticamības funkcijas definīcijā. Ja f ir blīvuma funkcija, tad šis saskaitāmais ir vienāds ar 0 un izteiksme (1.1.1) sakrīt ar parasto log-lokālas ticamības funkcijas definīciju. Iemesls šim saskaitāmajam ir tāds, ka $\mathcal{L}(f)$ var būt kā lokālā ticamības funkcija jebkurai ne-negatīvai funkcijai f bez ierobežojuma ($\int f(x) = 1$).

Log-lokālas ticamības funkcijas novērtējuma īpašība ir

$$E_f(\mathcal{L}(f_1)) \leq E_f(\mathcal{L}(f)). \tag{1.1.2}$$

Visām blīvuma funkcijām f_1; vienādība $f_1 = f$ izpildās gandrīz vienmēr. No (1.1.1) šī īpašība saglabājas jebkurām ne-negatīvajām, integrējamām funkcijām f_1 definētām kopā $\mathcal{X}$ un f_1 nav obligāti jābūt blīvuma funkcijai.

Lokālā log-ticamības funkcija

$$\mathcal{L}_x(f) = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right) log(f(X_j)) - n\int_{\chi} W\left(\frac{u - x}{h}\right) f(u)du, \tag{1.1.3}$$

kur W ir piemērota ne-negatīva svaru funkcija un h ir joslas platums. Mēs pieņemam, ka h ir konstante, ko praksē bieži vien tiek izvēlēta no datiem. Arī rezultātu var mainīt un uzlabot atkarībā no h izvēles, to mainot tāpat kā x vai X_j. Līdz ar to lokālās ticamības funkcijas īpašība

$$E_f(\mathcal{L}(f_1, x)) \leq E_f(\mathcal{L}(f, x)), \tag{1.1.4}$$

ar vienādību, kad $f(u) = f_1(u)$ pie $W((u-x)/h)$. Tas nodrošina $f(x)$ novērtēšanu maksimizējot izteiksmi (1.1.3) pēc piemērotas funkciju klases.

Lokālā polinoma aproksimācija pieņem, ka $logf(u)$ var būt labi aproksimēts pie mazas polinoma pakāpes p tuvā x apkārtnē.Tas ir

$$logf(u) \approx \langle a, A(u-x)\rangle,$$

piemēram, lokālā kvadrātiska polinoma gadījumā

$$A(u-x) = a_0 + a_1(u-x) + \frac{1}{2}a_2(u-x)^2.$$

Ar šo aproksimāciju mēs panākam, ka

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)\langle a, A(X_j - x)\rangle - n\int_{\chi} W\left(\frac{u-x}{h}\right)\exp(\langle a, A(u-x)\rangle)du. \quad (1.1.5)$$

Definīcija 1. Priekš fiksēta $x \in \mathcal{X}$ iegūst parametru $\hat{a} = (\hat{a}_0, ..., \hat{a}_p)^T$, kurš tiek maksimizēts no (1.1.5). Pēc lokālās ticamības funkcijas blīvuma novērtējums ir definēts kā

$$\hat{f}(x) = \exp(\langle \hat{a}, A(0)\rangle). \quad (1.1.6)$$

Lokālā parametra vektors $\hat{a}$ ir lokālās ticamības funkcijas sistēmas vienādojuma atrisinājums diferencējot (1.1.5) (ja (1.1.5) nevaram diferencēt vai x nepieder kopai $\mathcal{X}$, tad $\hat{f}(x) = 0$) un iegūst:

$$\frac{1}{n}\sum_{i=1}^{n} A(X_j - x)W\left(\frac{X_j - x}{h}\right) = \int_{\chi} A(u-x)W\left(\frac{u-x}{h}\right)\exp(\hat{a}, A(u-x)du.$$

Piemērs 1. Ja $p = 0$, tad $\langle a, A(X_j - x)\rangle = a_0$.

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)a_0 - n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)\exp(a_0)du$$

$$(\mathcal{L}_x(a))'_{a_0} = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right) - n\exp(\hat{a_0})\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)du = 0$$

$$n\exp(\hat{a_0})\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)du = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)$$

$$\exp(\hat{a_0}) = \frac{1}{n\int_{\mathcal{X}} W(\frac{u-x}{h})du}\sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right) =$$

un veicot dažus pārveidojumus

$$\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)du = |\frac{u-x}{h} = v, (u-x) = hv, du = hdv|$$

$$= \int_{\mathcal{X}} hW(v)dv = h\int_{\mathcal{X}} W(v)dv.$$

$$= \frac{1}{nh\int_{\mathcal{X}} W(v)dv}\sum_{i=1}^{n} w_j(x)$$

, kas ir blīvuma funkcijas novērtējums.

Piemērs 2. Ja $p = 1$, tad $\langle a, A(X_j - x)\rangle = a_0 + a_1(X_j - x)$;

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)(a_0 + a_1(X_j - x)) - n\int_{\mathcal{X}} W\left(\frac{u - x}{h}\right)\exp(a_0 + a_1(u - x))du$$

$$(\mathcal{L}_x(a))'_{a_0} = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right) - n\exp(\hat{a_0})\int_{\mathcal{X}} W\left(\frac{u - x}{h}\right)\exp(a_1(u - x))du = 0$$

$$\exp(\hat{a_0}) = \frac{\sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)}{n\int_{\mathcal{X}} W\left(\frac{u-x}{h}\right)\exp(a_1(u - x))du}$$

$$(\mathcal{L}_x(a))'_{a_1} = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)(X_j - x) - n\int_{\mathcal{X}} W\left(\frac{u - x}{h}\right)\exp(\hat{a_0})(u - x)\exp(\hat{a_1}(u - x))du = 0$$

$$n\int_{\mathcal{X}} W\left(\frac{u - x}{h}\right)\exp(\hat{a_0})(u - x)\exp(\hat{a_1}(u - x))du = \sum_{i=1}^{n} W\left(\frac{X_j - x}{h}\right)(X_j - x).$$

Taču šajā gadījumā varam redzēt, ka atrisinājumu analītiskā formā nevarēs iegūt.
Ja $\hat{a}$ eksistē, tad tam vajadzētu atrasties vaļējā kopā χ un tādēļ sistēma apmierina lokālās ticamības funkcijas vienādojumu:

$$\frac{1}{n}\sum_{i=1}^{n} A(\frac{X_j - x}{h})W(\frac{X_j - x}{h}) = \int_{\chi} A(\frac{u - x}{h})W(\frac{u - x}{h})exp(a, A(u - x)\rangle)du, \qquad (1.1.7)$$

kur $A(v) = (1 \quad v \quad v^2 \quad ... \quad v^p)$ vienas dimensijas gadījumā [4]. Kā iepriekš tika minēts, ka (1.1.5) nav risinājuma analītiskā formā, bet ir svarīgs jautājums par par eksistenci un unitāti.

1.1.1. Eksistence un unitāte

Lai C ir parametru telpa (atkarīgs no atbilstošā punkta x, svaru funkcijas W un lokālā polinoma pakāpes p):

$$C = \{(a = a_0, .., a_p) : \int_{\mathcal{X}} W\left(\frac{u - x}{h}\right)\ \exp(\langle a, A(u - x)\rangle)du < \infty\}. \qquad (1.1.8)$$

Daudzos gadījumos $\mathcal{X}$ ir atvērtā intervālā, piemēram, ja svaru funkcija W ir ierobežota un kompakta, tad $C = R^d$. Šajā gadījumā parametru vektors $\hat{a}$ (ja tāds eksistē) atrodas C

iekšienē un tas ir risinājums no lokālās ticamības funkcijas (1.1.5) blīvuma novērtējuma. No lokālās ticamības funkcijas Jakobiāns ir:

$$J(a) = -\int_{\chi} A(u-x)A(u-x)^T W(\frac{u-x}{h}) \exp(\langle a, A(u-x)\rangle)du.$$

Ne-negatīvai svaru funkcijai W tas ir definēts negatīvi. Tas norāda, ka lokālās ticamības funkcija ir ieliekta funkcija un novērtējums, ja tāds eksistē, ir unikāls.

Teorēma 1. *(Eksistence) Pieņem, ka parametru kopa (1.1.8) ir atvērta, tad lokālās ticamības funkcijas blīvuma novērtējums eksistē, ja vien nav tāds parametrs $a_0 \neq 0$, ka*

$$\langle a_0, A(u-x)\rangle = 0 \quad \forall \quad i : W\left(\frac{X_i - x}{h}\right) > 0$$

$$\langle a_0, A(u-x)\rangle \leq 0 \qquad \forall \quad u : W\left(\frac{u-x}{h}\right) > 0$$

Pierādījumu var skatīt [5].

1.1.2. Krosvalidācija

Lokālās ticamības funkcijas Krosvalidācija blīvuma funkcijas novētējumam:

$$LCV(\hat{f}) = \sum_{i=1}^{n} log(\hat{f_{-i}}(X_i) - n(\int_{\mathcal{X}} \hat{f}(u)du - 1),$$

kur $f_{-i}(X_i)$ nozīmē, ka blīvuma funkcija tiek novērtēta, kad X_i tiek izņemts no datu kopas.

Atšķirīga metode ir Squared Error Krosvalidācijas metode

$$\int_{-\infty}^{\infty} (\hat{f}(x) - f(x))^2 dx = \int_{-\infty}^{\infty} \hat{f}(x)^2 dx - 2\int_{-\infty}^{\infty} \hat{f}(x)f(x)dx + \int_{-\infty}^{\infty} f(x)^2 dx$$

Redzam, ka trešais objekts nav atkarīgs no $\hat{f}(x)$, bet pārējie divi tiek izteikti:

$$\int_{-\infty}^{\infty} \hat{f}(x)f(x)dx = E(\hat{f}(X)),$$

$$E(\hat{f}(X)) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i),$$

$$\int_{-\infty}^{\infty} \hat{f}(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f_{-i}}(X_i),$$

kura arī tiek saukta par Least Squerred Krosvalidāciju.

1.1.3. Blīvuma funkcijas novērtējums ar *Locfit* paketi programmā R

Piemērs 3. Tiek izmantoti dati (no "Old Faithful" (no Weisber(1985) un Scott(1992)), kuri satur 107 vulkānu ilgumu. *Locfit* komanda atbilst blīvuma funkcijas novērtējumam. Programmā R *Locfit* paketē tiek izmantota komponente *alpha*, kura norāda uz ticamības joslas platumu. Visbiežāk alpha ir starp 0 un 1 un tiek pielietots kā vektors no divām komponentēm, $alpha = c(\alpha_0, \alpha_1)$. $h(x)$ tiks aprēķināts kā:

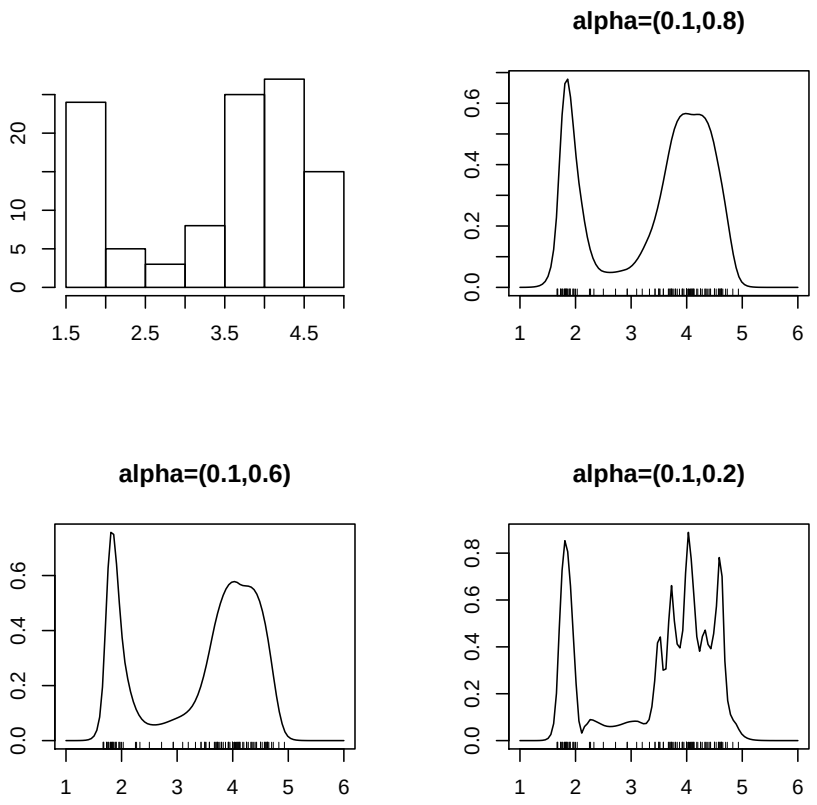
1. $k = \lfloor n\alpha_0 \rfloor$.

2. Rēķina $d_i = |x - x_i|$; $i = 1, ..., n$ un meklē mazāko d_k.

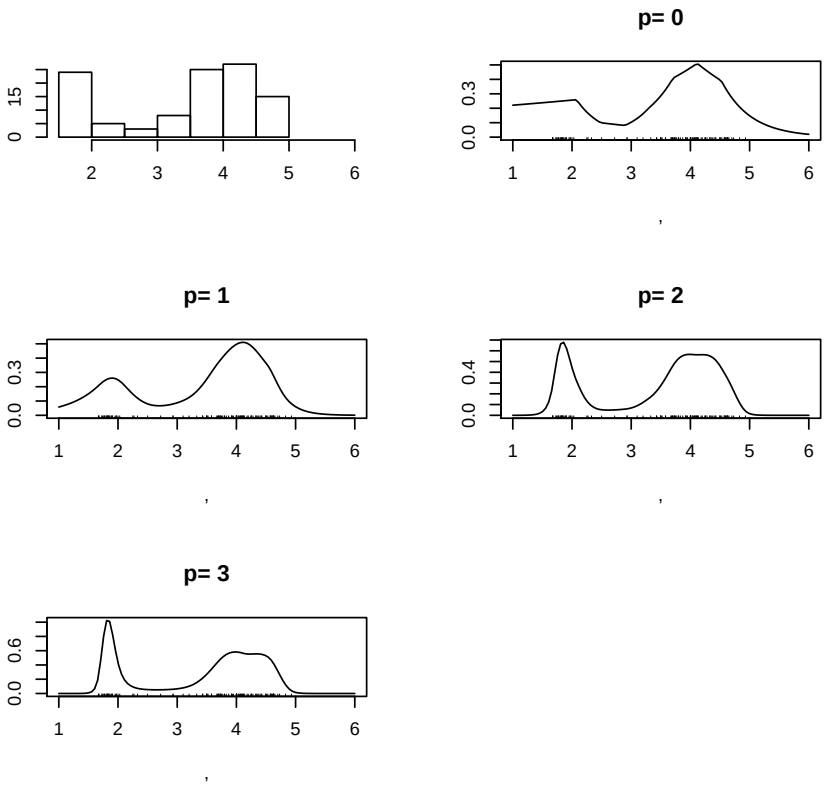
3. $h(x) = max(d_k, \alpha_1)$.

1.1. attēlā redzams, ka ir 2 smailes, kur pa kreisi smaile ir aptuveni pie 2. minūtēm un otrā smaile ir ap 4. minūtēm. Pie $alpha = c(0.1, 0.8)$, $alpha = c(0.1, 0.6)$ blīvuma funkcijas novērtējums ir labāks, taču pie $alpha = c(0.1, 0.2)$ tas ir pārgludināts.

1.2. attēlā blīvuma funkcijas novērtēšana tiek veikta ar jauktiem izlīdzināšanas parametriem - komponenti $alpha = c(0.1, 0.8)$ un pie dažādām polinoma pakāpēm. Kā labāko varētu atzīmēt lokālā kvadrātiskā un kubiskā polinoma pielietojumu.



1.1. att.: Blīvuma funkcijas novērtējums pie dažādiem gludināšanas parametriem alpha=c(0.1,0.8),alpha=c(0.1,0.6), alpha=c(0.1,0.2).



1.2. att.: Blīvuma funkcijas novērtējums pie dažādām polinomu pakāpēm, kad $p = 0$ (lokāls konstantes), $p = 1$(lokāls lineārs), $p = 2$ (lokāls kvadrātiskā), $p = 3$(lokāls kubiskā) polinoms.

2. Lokālās regresijas blīvuma novērtējums

2.1. Lokālā regresija

Lokālā regresija ir atkarība starp prediktoru X un mainīgo Y. Modeļa forma

$$Y_i = \mu(x_i) + \epsilon,$$

kur μ ir nezināmā funkcija un ϵ ir gadījuma kļūda un nav atkarīga no X_i vērtībām. Kļūda ϵ_i ir neatkarīga un vienādi sadalīta ar vidējo vērtību 0, $E(\epsilon_i) = 0$ un ar dispersiju $E(\epsilon^2) = \sigma^2 < \infty$. Tuvā punkta x apkārtnē mēs pieņemam, ka μ labi aproksimē parametrisku funkciju klasi. Piemēram, Teilora teorēma parāda, ka diferencējama funkcija var būt aproksimēta kā taisna līnija un divreiz diferencējamas funkcijas tiek aproksimētas kā kvadrātisks polinoms.

Izvēlētu punktu x definē ar joslas platumu $h(x)$ un gludināšanas intervālu $(x - h(x), x + h(x)$. Novērtē $\mu(x)$ tikai šī intervāla novērojumu robežās. Novērojumi tiek "svērti" pēc formulas:

$$w_i(x) = W\left(\frac{X_i - x}{h(x)}\right),$$

kur $W(u)$ ir svaru funkcija, kura x apkārtnē novērojumiem piešķir lielākus svarus. Piemēram, kuba svara funkcija

$$w(u) = (1 - |u|^3)^3$$

un Gausa svaru funkcija

$$W(u) = \exp(-(2.5x)^2/2).$$

Gludināšanas intervāla robežās $\mu(u)$ tiek aproksimēts kā polinoms. Piemēram, kvadrātiska polinoma aproksimācija ir

$$\mu(u) \approx a_0 + a_1(u - x) + \frac{1}{2}a_2(u - x)^2 \tag{2.1.1}$$

un $|u-x| < h(x)$. Īsāk varam pierakstīt, ieviešot apzīmējumu

$$a_0 + a_1(u-x) + \frac{1}{2}a_2(u-x)^2 = \langle a, A(u-x)\rangle,$$

kur a ir koeficientu vektors un $A(\cdot)$ ir vektors no dotas funkcijas. Priekš lokālās kvadrātiskas funkcijas

$$a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}; \quad A(v) = \begin{pmatrix} 1 \\ v \\ \frac{v^2}{2} \end{pmatrix}$$

Koeficientu vektors tiek novērtēts minimizējot kvadrātu summu:

$$\sum_{i=1}^{n} w_i(x)(Y_i - \langle a, A(X_i - x)\rangle)^2.$$

Un lokālās regresijas $\mu(x)$ novērtējums ir pirmā komponente $\hat{a}$.

Definīcija 2. Lokālās regresijas novērtējums konstanta polinoma gadījumā ir

$$\hat{\mu}(x) = \langle \hat{a}, A(0)\rangle = \hat{a}_0,$$

ko iegūst, ja $u = x$ ievieto (2.1.1).

2.1.1. Krosvalidācija

Krosvalidācijas ir metode kā mēs varam labāk izvēlēties blīvuma funkcijas novērtējumu. Sākumā šī metode tika pielietota parametriskam regresijas modelim, ko var skatīt (Allen(1974)), kas vēlāk arī tika pielietota splainu gludināšanai (Wahba nad Wold (1975)).

$$CV(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu_{-i}}(x_i)^2),$$

kur $\hat{\mu_{-i}}(x_)$ nozīmē, ka x_i tiek izmests no $\mu(x)$, t.i. katrs x_i tiek izmests no datu kopas un lokālās regresijas novērtējums tiek aprēķināts no $n-1$ datu punktiem.

2.2. Lokālās regresijas modelis

Lokālās ticamības funkcija regresijas modelim pieņem atbildes mainīgos kā blīvuma funkciju

$$(Y_i \sim f(y, \theta_i),$$

kur $\theta_i = \theta(x_i)$ un ir parametru vektors.

Lai $l(y, \theta) = log(f(y, \theta))$. Globālās ticamības funkcijas pielietojums regresijas modelim ir

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(Y_i, \theta(x_i)), \tag{2.2.1}$$

kur $\theta = (\theta(x_1), ..., \theta(x_n))$ ir parametru vektors.

Ģeneralizēts lineārais modelis pieņem $\theta(x)$ ir parametrisku lineāru forma, piemēram, $\theta(x) = a_0 + a_1 x$. Lokālā polinoma log-ticamības funkcija ir

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x) l(Y_i, \langle a, A(x_i - x)\rangle). \tag{2.2.2}$$

Veicot maksimizēšanu pēc parametra a iegūstam ticamības funkcijas novērtējumu.

Definīcija 3. (Lokālās ticamības funkcijas novērtējums) Maksimizē $\hat{a}$ no (2.1.2). Ticamības funkcijas novērtējums ir

$$\hat{\theta}(x) = \langle \hat{a}, A(0)\rangle = \hat{a}_0.$$

Piemērs 4. (Lokāls logistic regresijas modelis). Aplūko Bernulli regresijas modeli, kur

$$P(Y_i = 1) = p(x_i); \qquad P(Y_i = 0) = 1 - p(x_i)$$

Log-ticamības funkcija ir

$$\mathcal{L}(Y_i, p(x_i)) = logp(x_i)^{Y_i} + log(1 - p(x_i))^{1-Y_i} = Y_i log(p(x_i)) + (1 - Y_i) log(1 - p(x_i)) =$$

$$= Y_i(log(p(x_i)) - Y_i log(1 - p(x_i)) + log(1 - p(x_i)) =$$

$$= Y_i(p(x_i) - log(1 - p(x_i)) + log(1 - (p(x_i)) =$$

$$= Y_i log(\frac{p(x_i)}{1 - p(x_i)}) + log(1 - p(x_i)).$$

Lokālā polinoma aproksimācija varētu būt izmantota priekš $p(x_i)$. Bet tas nav vislabākais, tā kā $0 \leq p(x_i) \leq 1$, bet polinomam nav šādu ierobežojumu. Ja intervāls $(0, 1)$ tiek pārnests uz $(-\infty, \infty)$, tad izmanto logistic link funkciju

$$\theta(x) = log(\frac{p(x)}{1 - p(x)}).$$

Attiecīgi lokālā polinoma log-ticamības funkcija ir

$$\mathcal{L}_x(a) = \sum_{i=1}^{n} w_i(x)(Y_i\langle a, A(x_i - x)\rangle - log(1 + e^{\langle a, A(x_i - x)\rangle})).$$

Lokālā polinoma novērtējums ir $\hat{\theta}(x) = \hat{a}_0$. Novērtējot $p(x)$ iegūstam

$$\hat{p}(x) = \frac{e^{\hat{\theta}(x)}}{1 + e^{\hat{\theta}(x)}}.$$

Definīcija 4. (Link funkcija) Pieņem $f(y, \theta)$ ir sadalījuma parametriska ģimene ar vidējo

$$\mu = \mu(\theta) = E_\theta(Y)$$

un $\theta = g(\mu)$, kur g ir link funkcija. Inversais attēls, ir $\mu = g^{-1}\theta(x)$ un lokālās ticamības funkcijas novērtējums no $\mu(x)$ ir

$$\hat{\mu} = g^{-1}(\hat{\theta}(x)).$$

Parametriskā regresijas modeļa link funkcijas izvēle atkarīga no datiem. Ja vidējā vērtība ir log-lineāra, tad ir jāizmanto log-link funkcija. Lokālās regresijas modelim link funkcija tiek izvēlēta pēc izdevīguma. Taču svarīga prasība izmantojot logistic link funkciju ir parametra $\theta(x)$ intervāls $(-\infty, \infty)$. Priekš ne-negatīviem parametriem log link funkcija bieži piemērotākā izvēle. Otra prasība ir, lai $l(y, \theta)$ būtu ieliekts, kas garantē lokālās ticamības funkcijas pielietojuma stabilitāti regresijas modelim.
Dispersijas stabilizējošais link atbilst

$$-E\frac{\sigma^2}{d\theta^2}l(Y, \theta)$$

ir vienmerīgs, neatkarīgs no parametra θ. Ja link funkcija atbilst šīm īpatnībām, tad $var(\hat{\theta(x)})$ ir neatkarīgs no $\theta(x)$.
Eksponencionāla sadalījuma ģimenei blīvuma funkcija ir

$$f(y, \mu) = \exp(\tau(\mu)y - \psi(\mu))f_0(y)$$

Canonical link ir $\theta = \tau(\mu)$. Kad lokālā polinoma modelis ir izmantots priekš $\theta(x)$, lokālā ticamības funkcija (un tādēļ $\hat{\theta}(x)$) $\mathcal{L}_x(a)$ ir atkarīga tikai no datiem $\sum_{i=1}^{n} w_i(x)A(x_i - x)Y_i$. Tas vienkāršo teorētiskos apreķinus.
Kapēc tad maksimizējam lokālo ticamības funkciju?
Log-likelihood $\mathcal{L}(x)$ fiksētam θ ir gadījuma manīgais un atkarīgs no novērojumiem $Y_1, ..., Y_n$.

Vidējais $E(\mathcal{L}(\theta))$ ir funkcija no parametru vektora θ, kura tiek maksimizēta pēc parametra θ. Katram parametram θ^* pastāv sakarība, ka

$$E(\mathcal{L}(\theta^*)) < E(\mathcal{L}(\theta)).$$

Motivācija lokālā regresijas modeļa maksimizēšanai - parametra θ vērtību piemērota izvēle no datiem. Tādējādi starp parametru vektoriem mēs izvēlamies vienu parametru, kurš maksimizē empīrisko log-likelihood.

Tāpēc:

$$E\sum_{i=1}^{n} w_i(x)l(Y_i, \theta_i^*) \leqslant E\sum_{i=1}^{n} w_i(x)l(Y_i, \theta_i))$$

ar vienādību, ja tikai $\theta_i^* = \theta_i$ visiem i ar $w_i(x) > 0$. Lokālās regresijas blīvuma modeļa novērtējuma parametrus apskata no šādas formas $\theta_i^* = \langle a, (A(x_i - x))\rangle$, pēc kuriem arī tas tiek maksimizēts.

Pieņemot, ka lokālam regresijas modelim ir nosacījumi, tad parametru vektors $\hat{a}$ ir arī risinājums

$$\sum_{i=1}^{n} w_i(x)A(x_i - x)\partial(Y_i, \langle a, A(x_i - x)\rangle) = 0, \tag{2.2.3}$$

ko iegūst diferencējot (2.1.2). Pārsvarā lokālās regresijas vienādojumam (2.1.3) nav atklāta risinājuma, tāpēc bieži izmanto iteratīvās metodes, kuras apskata vai eksistē tāds $\hat{a}$ un vai $\hat{a}$ ir unikāls?

2.2.1. Krosvalidācija

Lai izvēlētos piemērotāko lokālās ticamības funkcijas modeli mums vajag izmantot krosvalidāciju

Definīcija 5. Lokālās ticamības funkcijas krosvalidācijas kritērijs balstās uz x_i "izmešanu" no $\theta_{-i}(x_i)$

$$LCV(\hat{\theta}) = \sum_{i=1}^{n} D(Y_i, \hat{\theta}_{-i}(x_i)) = C - 2\sum_{i=1}^{n} l(Y_i, \hat{\theta}_{-i}(x_i)),$$

kur C ir atkarīgs no Y_i novērojumiem, bet ne no novērtējuma $\hat{\theta}(x)$ un tātad arī to ticamības joslas vai lokālā polinoma pakāpes. D ir kā "zaudējumu" funkcija $D(Y, \hat{\theta}) = 2(supl(Y, \theta) - l(Y, \hat{\theta})))$ priekš novērojumu pāriem (x, Y).

2.2.2. Lokālās ticamības regresijas modelis ar $Locfit$ paketi

$Locfit$ pakete parāda lokālās ticamības regresijas modeli ar daudzām ģimenēm un linka funkcijām.Automātiski tiek pieņemta Gausa ģimene (ja netiek norādīta kāda cita).

Piemērs 5. Izmantosim datus *ethanol* programmā R. Atbildes mainīgais ir NOx un prediktors E, kas sastāv no 88 datiem, kur NOx ir piesārņotāju koncentrācija izplūdes gāzē un E ir gaisa un degvielas maisījumu attiecība. Izmanto komandu

$> -locfit(NOx\ E, data = ethanol, alpha = 0.5)$,

1) Pirmais arguments norāda uz atbildes mainīgo un prediktoru, otrais arguments norāda uz datu kopu. Trešais arguments precizē ticamības joslas platumu. Piemēram, $\alpha = 0.5$ nozīmē, ka ticamības josla "klāj" 50 % no dotajiem datiem. Ticamības joslas varbūt pievienotas ar komandu

$> plot(fit, band = global)$.

Arguments *global* nozīmē, ka šis ticamības intervāls ir konstruēts ar pieņēmumu, ka atlikumu dispersija ir konstante. Ja $band = local$, tad tiek veikts mēģinājums novērtēt dispersiju lokāli.

2.) Gludināšanas parametrs. Ticamības intervālu kontrolē α. Ja α ir uzdots kā viens skaitlis, tad tas reprezentē novērtējumu tuvākajā punkta apkārtnē.

Izmainot gludināšanas parametru, mēs novērtējam lokālās regresijas blīvuma funkciju ar dotajiem datiem pie dažādiem gludināšanas parametriem.

$> -alp < -c(0.8, 0.6, 0.4, 0.2)$

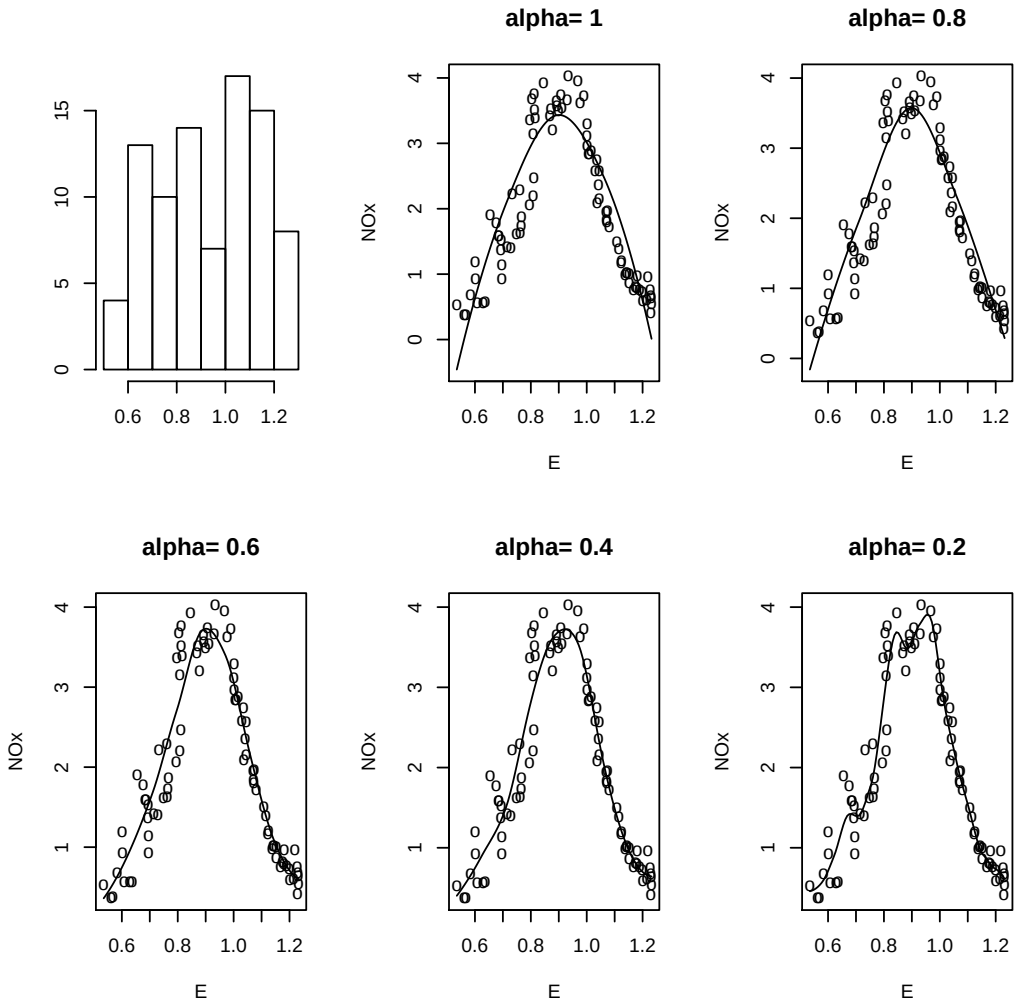
2.1.attēlā varam redzēt, ka pie $alpha = c(0.2)$ blīvuma funkcijas aproksimācija ir labāka nekā pie citiem gludināšanas parametriem. 3) *Alpha* varbūt izmantots kā vektors, kurš sastāv no divām komponentēm. Pirmā komponente ir fiksēts lielums, bet otrā komponenete reprezentē konstantu ticamības intervālu, ja $alpha = c(0, 1)$ nozīmē, ka $h(x) = 1$ ir izmantot visur. Ja abas no fiksētam tuvākās apkārtnes komponentēm ir nulles, tad $h(x)$ būs izvēlēts kā lielākā komponente.

4) Ja mēs gribam apskatīties lokālo polinomu ar dažādām pakāpēm, tad izmantojam komandu *deg*, kur $deg = 0$ (konstante), $deg = 1$ (lokāls lineārs), $deg = 2$ (lokāls kvadrātisks, kurš arī tiek izmantots pie noklusējuma).

$> -d < -0$

$> -fit < -locfit(NOx\ E, data = ethanol, alpha = c(0, 0.5), deg = d)$

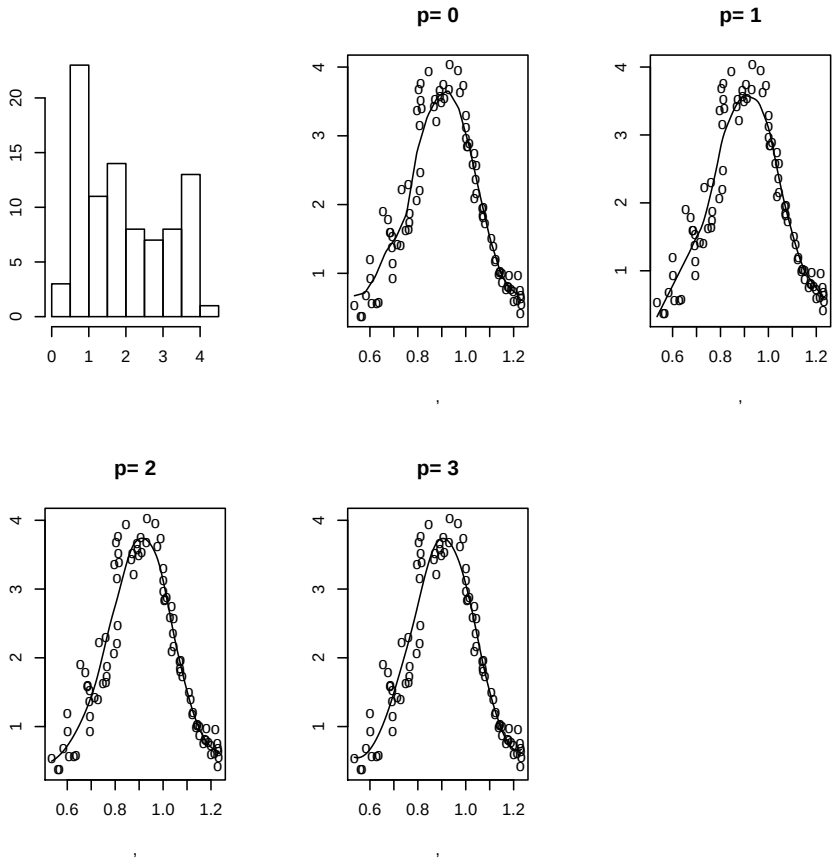
$> -plot(fit, get.data = T, main = paste("degree = ", d))$



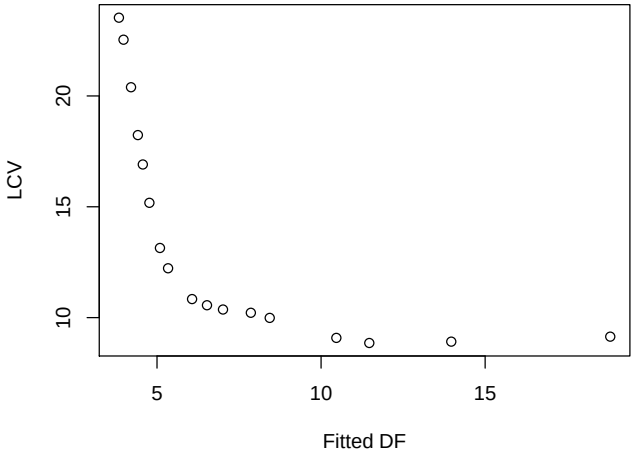
2.1. att. Lokālā regresijas modela novērtējums pie dažādiem gludināšanas parametriem.

Varam redzēt, ka 2.2.attēlā lokāls kvadrātisks polinoms un kubisks polinoms rāda labākus rezultātus, jo redzami mazāk trokšņu un arī dati tiek aproksimēti labāk nekā ar citām polinoma pakāpēm.

2.3. attēlā tiek veikta blīvuma funkcijas novērtējums ar Krosvalidācijas metodi.



2.2. att. Lokālās regresijas novērtējums pie dažādām polinoma pakāpēm.



2.3. att. Lokālā krosvalidācija.

Izmantotā literatūra un avoti

[1] Clive Loader. *Local Likelihood Estimation.* Department of Statistics, California, 1984.

[2] R. Tibshirani and T. Hastie. Local likelihood estiamtion. *Journal of American Statisical Assoiation*, 82:559–567, 1987.

[3] J.G. Staniswalis. The kernel estimate of a regression function in likelihood based models. *Journal of American Statisical Assoiation*, 84:276–283, 1989.

[4] R. Loader. Local likelihood density estiamtion. *The Analls of statistics*, 24:1602–1618, 1996.

[5] Clive Loader. *Local regression and likelihood.* Springer, USA, 1999.

Nobeigums

Savā kursa darbā izpētīju blīvuma funkcijas novērtējumu, izmantojot lokālo ticamības funkciju, kas no parastās ticamības funkcijas definīcijas atšķiras ar svaru pielietojumu, kas ļauj precīzāk veikt aproksimāciju. Sniegts neliels ieskats par lokālās ticamības funkcijas novērtējuma īpašībām, Krosvalidācijas metodi blīvuma un regresijas modeļa pielietojumam.

Praktiskā darba daļā esmu aplūkojusi blīvuma funkcijas novērtējumu un regresijas modeļa novērtējumu paketē $Locfit$ programmā R pie dažādiem joslas platumiem un polinoma pakāpēm p. Joslas platumu paketē $Locfit$ pielieto ar komandu $alpha$, kas cieši saistīts ar h. Izvēloties neprecīzu h var nonākt pie blīvuma funkcijas pārgludināšanas. Mainot polinoma pakāpes, tika apskatīts, ka labāks blīvuma funkcijas novērtējums ir pie lokālā kvadrātiskā un lokālā kubiskā polinoma izvēles.

Kursa darba uzrakstīšanai tika izmantota aprakstoša, salīdzinošā un analītiskā metode. Vēlētos vairāk izpētīt par labāku svaru funkcijas izvēli, kas atkarīgs no joslas platuma h. Darbu vēlētos turpināt, rakstot diplomdarbu.

Kursa darbs “Lokālā ticamības funkcija” 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: Oļesja Grigorčaka

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.

Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents:

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts kursa darbs gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)