Neparametriskās statistikas metodes ar pielietojumu laikrindu prognozēšanai

J. Valeinis[1]

[1]Latvijas Universitāte, Rīga

21.maijs, 2010

Neparametriskās un parametriskās metodes

Neparametriskās statistikas metodes:

- Blīvuma funkcijas novērtēšana ar kodolu metodēm;
- Regresijas funkcijas novērtējums ar kodolu metodēm un lokālā regresija;
- Butstrapa datu pārkārtošanas metodes;
- Empīriskā ticamības funkcija u.t.t.

Parametriskās metodes (pieņēmumi par populācijas sadalījumu):

- Vislielākās (maksimālās) ticamības funkcijas metode;
- Parametriskā regresija;
- t-tests u.t.t.

Histogramma

Doti $X_1, X_2, \ldots, X_n$ iid, kur $X_i \sim f$. Histogramma punktā x:

$$\hat{f}_n(x) = \frac{1}{2hn} \#\{X_i \in [x-h, x+h]\} = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

kur $K(u) = 0.5 1_{\{|u| \leq 1\}}$ ir vienmērīgā sadalījuma blīvuma funkcija (kodols) intervālā $[-1, 1]$.

- Histogramma ir neparametrisks blīvuma funkcijas novērtējums!
- Ideja: iegūt gludus (labākus) novērtējumus izvēloties citus (gludus) kodolus!

Neparametriskā blīvuma funkcijas novērtēšana

Kodolu neparametriskais blīvuma funkcijas novērtējums:

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

kur K-kodols, h-joslas platums.

- kodola izvēle K parasti nav būtiska, parasti izvēlas $N(0, 1)$ blīvuma funkciju (Gausa kodols);
- problēma: h izvēle!

Simulēti dati: *h* izvēle

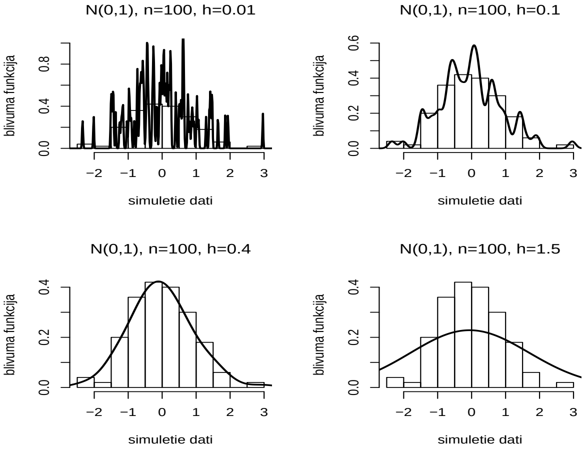

Figure: Kodolu gludināšana ar dažādiem *h*, kodols - Gausa

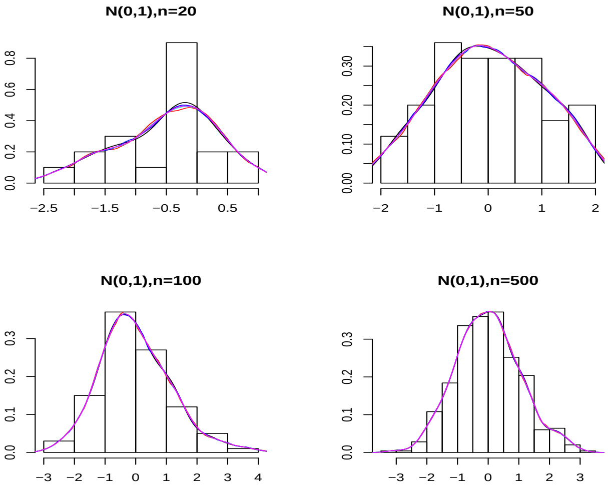

Figure: Kodolu gludināšana ar dažādiem kodoliem: "gaussian","biweight", "epanechnikov","rectangular","triangular","cosine", *h*-krosvalidācijas metode

Simulēti dati: histogramma & kodolu novērtējums

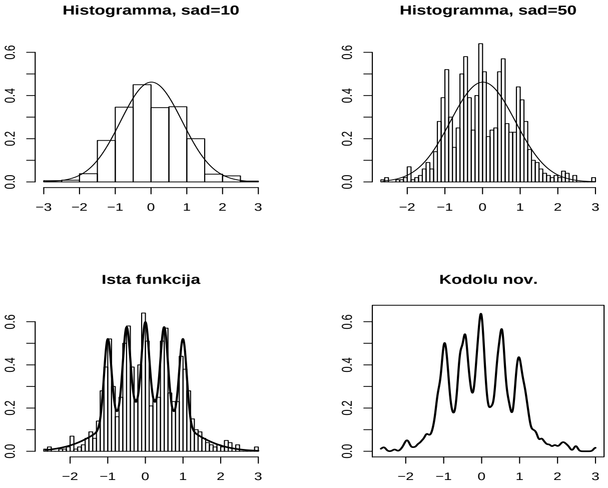

Figure: n=1000, p-vērtība KS-testam, Shapiro testam ir < 0.05

Joslas platuma izvēle

Vidējā kvadrātiskā kļūda (MSE) novērtējumam $\hat{f}_n(x)$:

$$MSE(\hat{f}_n(x)) = E((\hat{f}_n(x) - f(x))^2) =$$
$$= \frac{h^4}{4} f''(x)^2 \mu_2(K)^2 + \frac{1}{nh} ||K||_2^2 f(x) + o(h^4) + o\left(\frac{1}{nh}\right),$$

kur $\mu_2(K) = \int s^2 K(s) ds$ un $||K||_2^2 = \int K^2(s) ds$.

- Ideja: mizinimizēt integrētā vidējo kvadrātisko kļūdu: $\int MSE(f_n(x)) dx$.

- Optimālais h:

$$h_{opt} = \left(\frac{||K||_2^2}{||f''||_2^2 \{\mu_2(K)\}^2 n} \right)^{1/5} \sim n^{-1/5}.$$

- Problēma: h_{opt} satur nezināmo f''.
- "Rule of thumb": ja dati normāli sadalīti, tad $\|f''\|_2^2 = \sigma^{-5} \int \{\varphi''(x)\}^2 \, dx = \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0.212 \, \sigma^{-5}$. Tad

$$h_{opt} \approx 1.06 \hat{\sigma} n^{-1/5}.$$

Integrētā kvadrātiskā kļūda $ISE(h) = ISE(\hat{f}_n)$:

$$ISE(\hat{f}_n) = \int (\hat{f}_n(x) - f(x))^2 dx =$$
$$= ISE(h) = \int \widehat{f}_n^2(x)\, dx - 2 \int \{\widehat{f}_n f\}(x)\, dx + \int f^2(x)\, dx.$$

- Ievērosim, ka $\int \{\widehat{f}_n f\}(x)\; dx = E(\widehat{f}_n(X))$.
- Krosvalidācijas ideja: $\widehat{E\{\widehat{f}_h(X)\}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{f}_{h,-i}(X_i)$, kur

$$\widehat{f}_{h,-i}(x) = \frac{1}{(n-1)h} \sum_{j=1, i \neq j}^{n} K\left(\frac{x - X_j}{h}\right).$$

Lineārā (parametriskā) regresija

Doti datu pāri $(X_1, Y_1), ..., (X_n, Y_n)$. Regresijas vienādojums

$$Y_i = a + bX_i + \epsilon_i, \quad \mathbb{E}(\epsilon_i) = 0, \quad i = 1, ..., n,$$

Pieņēmumi: 1) ϵ_i ir neatkarīgi, vienādi sadalīti 2) $\epsilon_i \sim N(0, \sigma^2)$ (homoskedastisks modelis)

Polinomiālā regresija (ar pakāpi n):

$$Y_i = a_0 + a_1 X_i + a_2 X_i^2 + \ldots + a_n X_i^n + \epsilon_i.$$

Parametrus a un b novērtē pēc mazāko kvadrātu metodes

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 \rightarrow min!$$

Iegūst

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{a} = \bar{y} - \hat{\beta}\,\bar{x}.$$

Korelācijas koeficients

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{D(X)}\sqrt{D(Y)}}.$$

Īpašības

1. $-1 \leq \rho_{XY} \leq 1$;
2. Ja $Y = a + bX$, tad $\rho_{XY} = 1$ vai $\rho_{XY} = -1$;
3. Ja X un Y neatkarīgi, tad $\rho_{XY} = 0$;
4. $R^2 = \rho_{XY}^2$ raksturo, cik liela proporcija no Y datiem tiek izskaidrota ar X datiem.

Y - logaritms no divu lāzeru mērijumu attiecības; X - attālums.

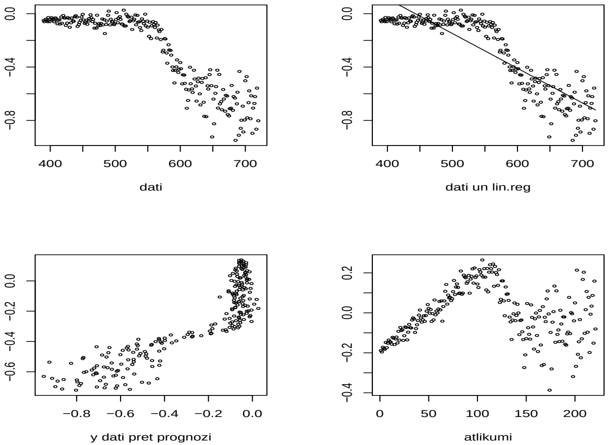

Figure: Lineārā regresija, n=221, $R^2 = 0.7827$, normalitāti nevar noraidīt, koeficienti ir nozīmīgi (tas ir var noraidīt $H_0 : a = 0$ un $H_0 : b = 0$)

LIDAR dati: lineārā regresija

Y - logaritms no divu lāzeru mērijumu attiecības; X - attālums.

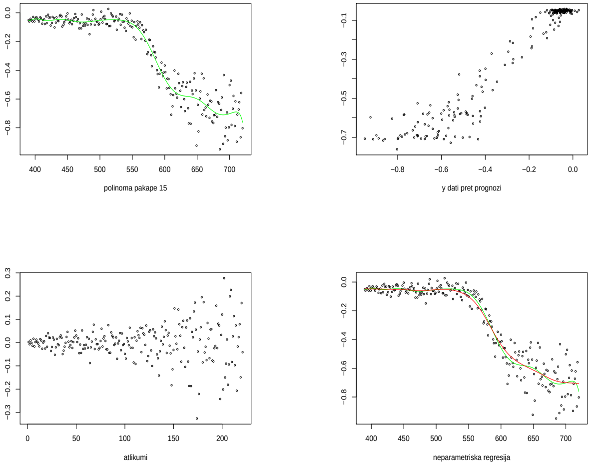

Figure: Polinomiālā (zaļa krāsa) un neparametriskā (sarkanā) regresija, n=221, $R^2 = 0.9253$, koeficienti ir nozīmīgi līdz 10 kārtai

Polinomu regresija: R izdruka

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.291156   0.005381 -54.108  < 2e-16 ***
poly(xx.data, 15)1  -3.706758   0.079994 -46.338  < 2e-16 ***
poly(xx.data, 15)2  -1.091555   0.079994 -13.645  < 2e-16 ***
poly(xx.data, 15)3   0.754951   0.079994   9.438  < 2e-16 ***
poly(xx.data, 15)4   0.617134   0.079994   7.715 5.20e-13 ***
poly(xx.data, 15)5  -0.254850   0.079994  -3.186  0.00167 **
poly(xx.data, 15)6  -0.369616   0.079994  -4.621 6.76e-06 ***
poly(xx.data, 15)7   0.135033   0.079994   1.688  0.09293 .
poly(xx.data, 15)8   0.246893   0.079994   3.086  0.00231 **
poly(xx.data, 15)9  -0.074803   0.079994  -0.935  0.35083
poly(xx.data, 15)10 -0.238201   0.079994  -2.978  0.00325 **
poly(xx.data, 15)11 -0.084672   0.079994  -1.058  0.29109
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07999 on 205 degrees of freedom
Multiple R-squared: 0.9253,     Adjusted R-squared: 0.9198
F-statistic: 169.2 on 15 and 205 DF,  p-value: < 2.2e-16
```

Doti datu pāri $(X_1, Y_1), ..., (X_n, Y_n)$. Regresijas vienādojums

$$Y_i = r(X_i) + \epsilon_i, \quad \mathbb{E}(\epsilon_i) = 0, \quad i = 1, ..., n,$$

kur $r(x) = \mathbb{E}(Y|X = x)$.

1. Nadaraya-Watson (1978) kodolu novērtējums

$$\hat{r}(x) = \frac{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right)},$$

kur K ir kodols (blīvuma funkcija) un h - joslas platums.

2. Lokālais lineārais regresijas novērtējums: ideja minimizēt

$$\sum_{i=1}^{n}(Y_i - a - b(X_i - x))^2 K\left(\frac{X_i - x}{h}\right)$$

Rezultāts:

$$\widehat{r}_n(x) = \frac{\sum_{i=1}^{n} b_i(x) Y_i}{\sum_{j=1}^{n} b_j(x)},$$

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

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

Lokālais lineārais novērtējums uzlabo robežu novirzi kodolu novērtējumam (svarīgi prognozēšanai)

Neparametriskā (kodolu) regresija: SP500 index

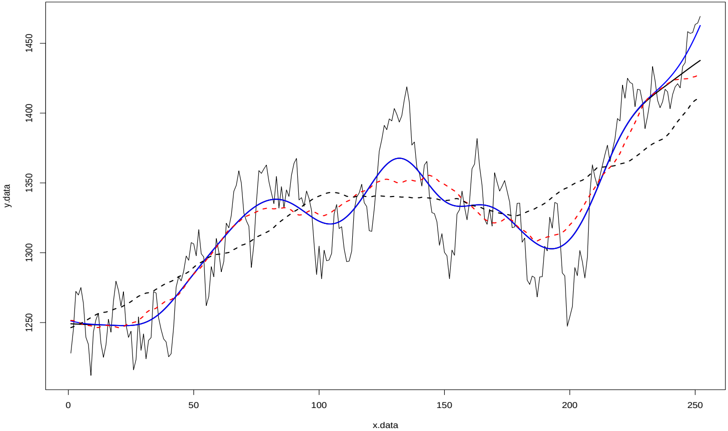

Figure: Neparametriskie regresijas novērtējumi (zilā svītra - lokālais lineārais nov.; melnā svītra - kodolu nov.; sarkanā - slīdošais vidējais (21 dienu intervāls); melnā raustītā - slīdošais vidējais (41 dienu intervāls)).

Neparametriskā (kodolu) regresija: problemātika

- Joslas platuma noteikšana atkarīgiem datiem (laikrindām, ARIMA modeļiem, jauktiem procesiem). Parastās metodes (krosvalidācija utt.) īsti nestrādā.
- Prognozes veikšana ar neparametrisko regresiju.
- Citi neparametriski gludinātāji: splaini, Veivletu regresija, utt.
- Butstrapa metodes neparametriskajā regresijā.

- Maģistra darbi: Haralds Plivčs (2009), Natālija Saveļjeva (2009)