Butstrapa metodes atkarīgiem datiem

Mārcis Bratka

2010

Efron (1979) noformulēja butstrapa metodi, kura ļauj novērtēt parametru sadalījumus ar vienkāršiem pieņēmumiem.

- $X_1, X_2, \ldots$ ir neatkarīgi un vienādi sadalīti gadījuma lielumi ar sadalījumu F. $\mathcal{X}_n = (X_1, \ldots, X_n)$ ir dotā izlase un $T_n = t_n(\mathcal{X}_n, F)$ ir funkcionālis, $G_n(x) \equiv P(T_n \leq x)$.
- Pārkārtojam elementus jaunā izlasē $\mathcal{X}_m^* = (X_1^*, \ldots, X_m^*)$ ar atkārtojumiem, kur $m \leq n$ (parasti $m = n$).
- X_i^* ir neatkarīgi un vienādi sadalīti ar $P_*(X_1^* = X_j) = n^{-1},\ 1 \leq j \leq n$.
- X_i^* sadalījums $F_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, kur δ_y apzīmē Diraka varbūtības mēru ar vērtību 1 punktā y un 0 citur.
- Butstrapa versija $T_{m,n}^* = t_m(\mathcal{X}_m^*, F_n)$ un tā sadalījums $\hat{G}_{m,n}$.

Statistika $T_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma$ un butstrapa versija $T^*_{m,n} = \sqrt{m}(\bar{X}^*_m - \bar{X}_n)/s_n$.

Teorēma

Ja $X_1, X_2, \ldots$ ir neatkarīgi un vienādi sadalīti, $EX_1^2 < \infty$, $\sigma^2 = DX_1 \in (0, \infty)$, tad $\sup_x |P_*(T^*_{n,n} \le x) - \Phi(x)| = o(1)$ gandrīz droši, kad $n \to \infty$, kur $\Phi(\cdot)$ ir $N(0,1)$ sadalījuma funkcija.

- $\sup_x |P_*(T^*_{n,n} \le x) - P(T_n \le x)| = o(1)$.
- Singh (1981) parādīja, ka pie dažiem nosacījumiem butstrapa metode dod aproksimāciju $\sup_x |P_*(T^*_{n,n} \le x) - P(T_n \le x)| = O(n^{-1}(\log\log n)^{1/2})$. Klasiskā aproksimācija dod $O(n^{-1/2})$.

Ja $\{X_n\}_{n\geq 1}$ ir atkarīgu gadījumu lielumu virkne, tad

$$\sigma^2 = DX_1 + \sum_i \mathrm{Cov}(X_1, X_{1+i}).$$

Pārkārtojot elementus nejaušā secībā tiek pazaudēta datu atkarības struktūra. Singh (1981) ar piemēru pierādīja, ka, ja $T^* = \sqrt{n}(\bar{X}^* - \bar{X})$, tad

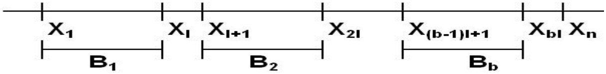
$$\lim_{n\to\infty} |P_*(T^* \leq x) - P(T \leq x)| \neq 0.$$

Butstrapa sadalījums joprojām tiecas uz normālo sadalījumu, taču ar nepareizu dispersiju.

Kunsch (1989) un Liu un Singh (1992) definēja butstrapa metodes, kuras ir lietojamas atkarīgiem datiem. Metožu galvenā ideja ir pārkārtot datu blokus.

- Slīdošo bloku butstraps (MBB)

- Nešķeļošo bloku butstraps (NBB)

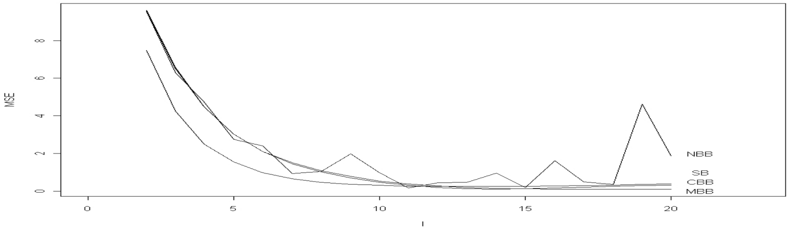


- Riņķveida bloku butstraps (CBB)

- Stacionārais butstraps (SB)

ARMA(1,1) modelis $X_i = 0.2X_{i-1} + \epsilon_i - 0.4\epsilon_{i-1}$ ar $T_n^* = \sqrt{n}(\bar{X}_n^* - \mu)$ sadalījumu $N(0, 0.645)$.

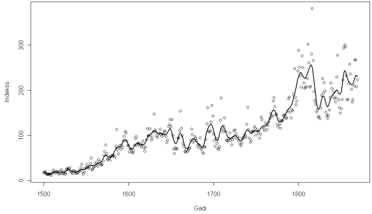
	iid butstraps		Bloku butstraps	
Izlases apjoms	Vidējā vērtība	Dispersija	Vidējā vērtība	Dispersija
50	-0.003	1.334	-0.003	0.544
100	-0.016	0.966	0.042	0.358
200	0.024	0.913	0.014	0.389
500	-0.019	1.087	0.003	0.625
1 000	-0.033	1.035	0.040	0.637
2 000	0.016	0.972	0.004	0.672
5 000	-0.028	0.965	0.013	0.679



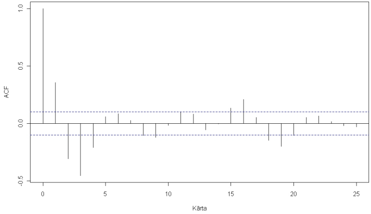
- $l_{opt} = \arg\min \mathrm{MSE}(l)$,
- MBB un CBB metodes ir precīzākās MSE nozīmē,
- MBB aproksimācija dod precīzākus novērtējumus parametriem $\theta = H(\mu)$ kā normālā aproksimācija, kur H ir gluda funkcija.

Neparametriskās regresijas lietojumos ir nepieciešams notiekt joslas platumu. Viena no daudzām metodēm ir joslas platuma iegūšana ar butstrapa metodi:

- sākuma h_0,
- regresijas novērtējums $\hat{m}_{h_0}(x)$,
- regresijas atlikumi $\epsilon_i = Y_i - \hat{m}_{h_0}(X_i)$,
- centrēti un normēti atlikumi $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$,
- butstrapa izlase $\epsilon_1^*, \ldots, \epsilon_n^*$,
- butstrapa izlase $Y_i^* = \hat{m}_{h_0}(X_i) + \epsilon_i^*$,
- MISE novērtējums $\widehat{\mathrm{MISE}}(h) = (nB)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{B} \left(\hat{m}_{j,h}^*(X_i) - \hat{m}_{h_0}(X_i) \right)$,
- optimālais joslas platums $h_{opt} = \text{minarg}\{\widehat{\mathrm{MISE}}(h)\}$,
- atkārtojam soli 1-8 $h_0 = h_{opt}$.

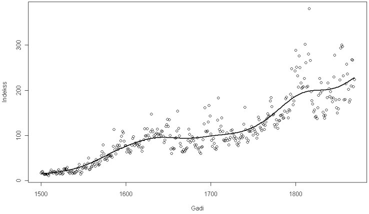


(a)

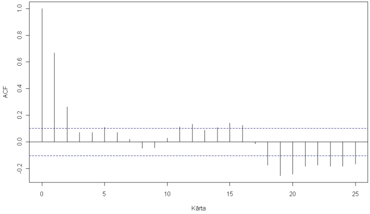


(b)

CV, AICC un butstrapa metodes nestrādā gadījumos, kad regresijas atlikumi ir atkarīgi.
Piemērs ar Beveridžas indeksu datiem. Joslas platums iegūts ar CV (a). Regresijas atlikumu analīze (b) uzrāda korelācijas pazīmi.



(c)



(d)

Gadījumos, kad regresijas atlikumi ir atkarīgi, ir jālieto bloku butstrapa metode joslas platuma izvēlei. Atlikumi tiek pārkārtoti sadalīti blokos.

- Matched block bootstrap
  $B_{i_1}, \ldots, B_{i_j}$, $p(i_j, i_{j+1})$ ir varbūtība, ka $j+1$ bloks ir $B_{i_{j+1}}$.
- Tapered block bootstrap
  Viedējās vērtības gadījums:
  $X^*_{ml+j} = w_l(j)\frac{l^{1/2}}{\|w_l\|_2}(X_{i_m+j-1} - \bar{X})$, $j = 1, \ldots, l$, kur $w_n(t) = w((t-0.5)/n)$ un optimālā $w_c^{trap}(t) = (t/c)1(t \in [0,c)) + 1(t \in [c, 1-c]) + ((1-t)/c)1(t \in (1-c, 1])$ un $c = 0.43$.
  - Tapered block bootstrap dod precīzākus novērējumus, kā MBB MSE nozīmē (samazina novirzes daļu).

- Model based bootstrap
  $X_i = \beta_1 X_{i-1} + \ldots + \beta_p X_{i-p} + \epsilon_i,\ i \in Z.$
  - Līdzīga pieeja kā regresijas gadījumā.
  - Ja ir izvēlēts pareizs modelis, tad precizitāte ir augstāka kā bloku butstrapa metodēm.
- Frequency domain bootstrap

- Dependent wild bootstrap
  $X_i^* = \bar{X}_n + (X_i - \bar{X}_n)W_i$, kur $\{W_i\}_{i=1}^n$ ir gadījuma lielumi neatkarīgi no X_i. $E(W_i) = 0$, $Var(W_i) = 1$, $i = 1, \ldots, n$ un $Cov(W_i, W_i') = a((i - i')/l_n)$, kur $a(\cdot)$ ir kodola funkcija un l_n ir joslas platums. Piemērs, $\{W_{t_j}\}_{j=1}^n \sim N(0, \sum_W)$, kur $\sum_W = [a(t_i - t_j)/l_n]$, $i, j = 1, \ldots, n$.
  - Dependent wild bootstrap precizitāte vidējās vērtības gadījumā ir vienāda ar Tapered block bootstrap.
  - Dependent wild bootstrap darbojas gadījumā, kad dati ir neregulāri sadalīti.
  - Parametrs l_n var nebūt vesels skaitlis.
  - Šobrīd nav informācijas par second order accuracy.