BOOTSTRAP

STATISTIČKI PRAKTIKUM 2

10. VJEŽBE

Bootstrap metode su neparametarske motode ponovnog uzorkovanja korištene za procjenu pouzdanosti modela i procedura. Odlike:

- brza i jednostavna procjena
- bez pretpostavki na tip modela (ukoliko je model nepoznat ili kompleksan)
- ne oslanja se na asimptotske rezultate

Često se koriste i za analizu standardnih procjena u parametarskim modelima te za poboljšanje istih.

Osnovna ideja

Neka je $x_1, \ldots, x_n$ realizacija slučajnog uzorka $X_1, \ldots, X_n$ iz nepoznate distribucije F. Osnovna ideja je kako realizacije sadrže sve dostupne informacije o distribuciji F i stoga očekujemo da će ponovno uzorkovanje iz tog uzorka dati uzorak koji odgovara uzorkovanju iz distribucije F.
Promatrajmo parametar θ iz distribucije F i njegov procjenitelj $\hat{\theta} = S(X_1, \ldots, X_n)$. Često nas zanima i distribucija tog procjenitelja (npr. zbog pouzdanih intervala ili testiranja), ali nju je rijetko moguće odrediti.
Kada bi F bila poznata mogli bismo odrediti uzoračku distribuciju od $\hat{\theta}$ opetovanim uzorkovanjem iz F (Monte Carlo metoda).
Budući da je F nepoznata, uzorkujemo iz empirijske distribucije $\hat{F}$ dobivene na temelju realizacije $x_1, \ldots, x_n$.

Zadatak 1

Neka je $X_1, ..., X_n$ uzorak iz normalne razdiobe $N(\mu, \sigma^2)$. Poznato je da je

$$H = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1).$$

Provjerite ovu tvrdnju Monte Carlo simulacijama za $n = 100$, $\mu = 1$ i $\sigma^2 = 4$. Broj simulacija $R = 50$.

Uzorkovanje iz $\hat{F}$

Želimo uzorak $x_1^*, \ldots, x_k^*$ duljine k iz $\hat{F}$,

1. uzorkujemo $i_1, \ldots, i_k$ iz uniformne na $\{1, \ldots, n\}$,
2. definiramo $x_j^* = x_{i_j}$.

Na temelju tog uzorka možemo odrediti bootstrap procjenitelj za θ

$$\hat{\theta}^* = S(X_1^*, \ldots, X_k^*),$$

$$\hat{\mathbb{P}}(\hat{\theta} \in A) = \mathbb{P}^*(\hat{\theta}^* \in A),$$

gdje potonju distribuciju možemo dobiti Monte Carlo metodama iz $\hat{F}$.

Zadatak 2

U datoteci `sample.txt` dan je uzorak duljine 1000 iz nepoznate razdiobe s konačnom varijancom. Bootstrap metodom provjerite da ta razdioba zadovoljava centralni granični teorem.

Greške pri procjeni

Postoje dva izvora greške:

- zamjena F sa $\hat{F}$
- procjena distribucije od $\hat{\theta}$ Monte Carlo simulacijama iz $\hat{F}$

Druga greška može se proizvoljno smanjiti, relativno na prvu, i to odabirom većeg broja uzoraka u MC simulaciji.

Pouzdani intervali za θ

Postoji nekoliko mogućih pristupa za bootstrap procjenu pouzdanih intervala

1. normalna aproksimacija
2. osnovni bootstrap
3. studentizirani bootstrap
4. percentile
5. BC (*bias-corrected*)

Normalna aproksimacija

Pretpostavimo da je

$$Z = \frac{\hat{\theta} - \theta - b(F)}{\sigma(F)} \sim N(0,1),$$

gdje je $b(F) = \mathbb{E}(\hat{\theta}|F) - \theta$ i $\sigma^2(F) = Var(\hat{\theta}|F)$. Korištenjem bootstrap metode procijenimo nepoznate $b(F)$ i $\sigma(F)$ s

$$b(\hat{F}) = \frac{1}{R}\sum_{j=1}^{R}\theta_j^* - \theta(\hat{F})$$

$$\sigma^2(\hat{F}) = \frac{1}{R}\sum_{j=1}^{R}(\theta_j^* - \bar{\theta}^*)^2, \qquad (1)$$

gdje je R broj MC simulacija iz $\hat{F}$, θ_j^* procjenitelj za θ u j-toj simulaciji, $\theta(\hat{F})$ procjena parametra θ na temelju $\hat{F}$.

Osnovni i percentile bootstrap

Kvantili distribucije od $\hat{\theta} - \theta$ su približno jednaki kvantilima distribucije od $\bar{\theta}^* - \theta(\hat{F})$, pa je pogodan pouzdani interval za $\hat{\theta}$

$$\left[2\theta(\hat{F}) - \theta^*_{((R+1)(1-\alpha/2))}, 2\theta(\hat{F}) - \theta^*_{((R+1)\alpha/2)}\right].$$

Druga mogućnost je da za pouzdani interval uzmemo

$$\left[\theta^*_{((R+1)\alpha/2)}, \theta^*_{((R+1)(1-\alpha/2))}\right].$$

Bootstrap u R-u

```
> install.packages("boot")
> library(boot)

> boot
function (data, statistic, R, sim = "ordinary", ...)

> boot.ci
function (boot.out, conf = 0.95, type = "all", ...)
```

Primjer

U datoteci `bodovi.txt` nalaze se bodovi 200 ispitanika iz provjere pismenosti i matematičkog dijela testa. Odredite 90% pouzdani interval za koeficijent korelacije uspjeha u tim područjima.
$\theta = \rho(write, math)$, $\hat{\theta} = cor(write, math)$

```
> cor(bodovi$write, bodovi$math)
[1] 0.6174493
> corr <- function(d, i){
+ d2 = d[i,]
+ return(cor(d2$write, d2$math))
+ }
> bootcorr = boot(bodovi, corr, R=500)
```

```
> bootcorr

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = bodovi, statistic = corr, R = 500)

Bootstrap Statistics :
     original         bias    std. error
t1* 0.6174493 -6.345684e-05  0.04048737

> names(bootcorr)
> plot(bootcorr)
```

```
> boot.ci(bootcorr, conf=0.9)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 500 bootstrap replicates

CALL :
boot.ci(boot.out = bootcorr, conf = 0.9)

Intervals :
Level      Normal              Basic
90%   ( 0.5509,  0.6841 )   ( 0.5540,  0.6832 )

Level     Percentile            BCa
90%   ( 0.5517,  0.6809 )   ( 0.5469,  0.6775 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(bootcorr, type = "all") :
  bootstrap variances needed for studentized intervals
```

Zadatak

Simulirajte uzorak duljine 1000 iz standardne normalne razdiobe. Bez pretpostavki na distribuciju iz koje podaci dolaze procijenite 95% pouzdani interval za očekivanje i medijan populacije.

(a) Koristite bootstrap procedure implementirane u `R`-u.

(b) Proceduru za *percentile* pouzdani interval provedite sami.