Računarska statistika

Snježana Lubura Strunjak

Zagreb, 20. svibnja 2021.

Metode ponovnog uzorkovanja (re-sampling)

Bootstrap metoda

Bootstrap metodu je uveo Bradley Efron (Stanford University) 1979. godine.
Efronovim riječima:
Statistika se provodi na 2 razine:

- Algoritmi (procjene)
- Određivanje točnosti tih procjena

Bootstrap je način na koji se, uz pomoć računala odgovara na pitanje o točnosti procjene (jer su algoritmi postali kompleksni, a količina podataka enormna).
Pročitati: lect13.pdf u PDF Materijali na Merlinu.

Bootstrap pristupi

- Parametarski bootstrap: Simulacija podataka po procijenjenom (fitted) modelu:
  1. Pretpostavite parametarski model za podatke
  2. Fitajte model na podatke i procijenite parametre
  3. Uzorkujte iz procijenjenog (fitted) modela (ili distribucije)
- Neparametarski bootstrap: Slučajno uzorkovanje iz podataka tj. uzimanje slučajnih uzoraka s ponavljanjem (eng. with replacement) iz stvarnih podataka

ZAŠTO?
Za odredjivanje točnosti procjena:

- Procjenjivanjem pristranosti, varijance, intervala pouzdanosti
- Od posebnog značaja u situacijama gdje analitičkih formula nema ili nisu pouzdane.

Neparametarski bootstrap

Primjer (Neparametarski bootstrap za sredinu)

Neka je dan uzorak veličine 6: $3.12, 0, 1.57, 9.67, 0.22, 2.22$

Parametar (θ) je očekivanje kojeg procjenjujemo s prosječnom vrijednosti/sredinom.

Procjena na uzorku je $\hat{\theta} = 2.8$ (aritmetička sredina uzorka).

Na slučajan način birajmo uzorke duljine 6 s ponavljanjem iz danog uzorka (to sve ponavljamo B puta), i za svaki od tih novih uzoraka izračunajmo procjenu parametra , odnosno aritmetičku sredinu, u oznaci $\hat{\theta}^\star(i)$, $i = 1, \ldots, B$.

U našem primjeru:

- 1.uzorak: 0,0,0,1.57,9.67,0.22, $\hat{\theta}^\star(1) = 1.91$
- 2.uzorak: 3.12, 9.67, 0.22, 0.22,2.22,2.22, $\hat{\theta}^\star(2) = 2.94$
- ...

Bootstrap procjena očekivanja (sredine) je: $\frac{\sum_{i=1}^{B} \hat{\theta}^\star(i)}{B} = 2.88$, procjena standardne devijacije je $\hat{\sigma}_B = 1.35$ (ovdje je B u nazivniku).

Procjena pristranosti pomoću bootstrapa je: $2.88 - 2.8 = 0.08$

90% CI (engl. confidence interval, odnosno pouzdani interval) se dobije percentilnom metodom: $[1.0, 5.4]$.

Primjer (Neparametarski bootstrap za medijan)

Neka je dan uzorak veličine 6: $3.12, 0, 1.57, 9.67, 0.22, 2.22$

Parametar (θ) je medijan. Procjena na uzorku je $\hat{\theta} = 1.895$.

Na slučajan način birajmo uzorke duljine 6 s ponavljanjem iz danog uzorka (to sve ponavljamo B puta), i za svaki od tih novih uzoraka izračunajmo procjenu parametra , odnosno medijan, u oznaci $\hat{\theta}^\star(i)$, $i = 1, \ldots, B$.

U našem primjeru:

- 1.uzorak: 0,0,0,1.57,9.67,0.22, $\hat{\theta}^\star(1) = 0.11$
- 2.uzorak: 3.12, 9.67, 0.22, 0.22,2.22,2.22, $\hat{\theta}^\star(2) = 2.22$
- ...

Bootstrap procjena medijana je: $\frac{\sum_{i=1}^{B} \hat{\theta}^\star(i)}{B} = 2.07$, procjena standardne devijacije je $\hat{\sigma}_B = 1.5$ (ovdje je B u nazivniku).

Procjena pristranosti pomoću bootstrapa je: $2.07 - 1.895 = 0.175$

90% CI (engl. confidence interval, odnosno pouzdani interval) se dobije percentilnom metodom: $[1.0, 5.4]$.

Bootstrap - ideja

Neka je $X_1, \ldots, X_n$ slučajni uzorak iz funkcije gustoće f, i neka je $\hat{f}_n$ histogram uzorka.
Neka je θ nepoznati parametar gustoće f, i neka je $\hat{\theta}$ procjena tog parametra na osnovu uzorka (uočimo da ta procjena ovisi o f).
Kako $\hat{f}_n$ aproksimira funkciju f, to će onda distribucija $\hat{\theta}^\star(1), \ldots, \hat{\theta}^\star(B)$ (bazirana na $\hat{f}_n$) aproksimirati distribuciju $\hat{\theta}$ (baziranu na f), odnosno sampling distribuciju.

Neki poznati bootstrap rezultati

- Aproksimacija sampling distribucije i intervala pouzdanosti je bolja što je n veći
- Bootstrap 90% percentilni interval pouzdanosti (CI) se dobije pomoću 5percentila i 95percentila bootstrap distribucije. Za njegovu preciznost je potrebno da je $B \geq 1000$.
- S porastom n bootstrap nikad nije lošiji od aproksimacije s centralnim graničkim teoremom, a za puno parametara može biti čak i puno bolji.

Kada ne koristiti percentilne bootstrap intervale pouzdanosti:

- Kada je bootstrap distribucija jako nesimetrična (npr. kod kvantila, oprez ako je n mali)
- Kada je statistika jako pristrana (velika razlika između bootstrap sredine i $\hat{\theta}$)
- Ako je potrebna velika preciznost
- Upozorenje: bootstrap nije magično rješenje za loše podatke (neprezentativnost, mali uzorak i slično)

Bolji bootstrap intervali pouzdanosti

- BC (bias corrected)
- BCa (bias corrected and accelerated)
  - Prilagodba percentilnih intervala tako da se korigira pristranost i asimetrija
  - Precizni rezultati u velikom broju situacija
- Bootstrap t
- Hall-ovi percentilni intervali
- Dvostruki bootstrap intevali

Bootstrap u SAS-u

Neparametarski bootstrap

- Jackboot macro
  - Složenija primjena
  - Razni int. pouzdanosti
- Procedura SURVEYSELECT
  - Jednostavnija primjena
  - Samo percentilni int.pouzdanosti
- U nekim je procedurama bootstrap ugrađen(NLIN, TPSPLINE, UCM, MULTTEST, QUANTREG, itd.)
- Data step

Primjer (Neparametarske bootstrap procjene za varijancu i medijan pomoću jackboot makro)

Program *CHAPTER*1_3_*bootstrap var i median.SAS* i data set LAW i MEDIAN (folder Data na Merlinu).

Program *jackboot.sas* ili otvoriti i pokrenuti, ili definirati path u SAS-u pa koristiti naredbu include:

%*let path* = /*home*/*u*47429085; i

%*include* "&*path*/*jackboot.sas*";

Nađimo neparametarsku bootstrap procjenu varijance koristeći jackboot makro, te zatim

- U pozivu %*boot*(*data* = *lib.law*) dodajte opciju *samples* = 1000, kako biste promijenili vrijednost B na 1000.
- Ispišite percentilne bootstrap i bootstrap BC intervale (%*bootci*(*PCTL*) i %*bootci*(*BC*)).
- Nacrtajte histogram (koristeći Tasks and Utilities) za podatke *var_LSAT* u data setu BOOTDIST. Uočite da je to bootstrap distribucija varijance koja aproksimira sampling distribuciju.

Nađimo neparametarsku bootstrap procjenu medijana koristeći jackboot makro, te zatim ispišite percentilne bootstrap i bootstrap BC intervale (%*bootci*(*PCTL*) i %*bootci*(*BC*)).

Bootstrap unutar Data step-a

Primjer (Bootstrap procjene koef. asimetrije i spljoštenosti (za Fisherove iris podatke))
Program *CHAPTER*1_3_*bootstrap skewness and kurtosis*.*sas*. Podaci su sadržani unutar SAS-a (sashelp.iris) ili u data set-u IRIS u folderu Data na Merlinu.
Uočite da u programu postoje dva načina uzorkovanja s ponavljanjem unutar Data step-a (prvi je unutar linija koda 32-47, a drugi unutar linija koda 55-72).

Neparametarski bootstrap s SERVEYSELECT naredbom

Primjer

Program unutar foldera Primjeri
Primjer Neparametarski Bootstrap sa surveyselect verzija44.SAS.
Usporediti s odgovarajućim dijelom programa
Primjeri – objasnjenje JACKBOOT macroa.SAS u folderu Primjeri na Merlinu.

Parametarski bootstrap

- Kod parametarskog bootstrapa pretpostavljamo neki odredjeni parametarski model za F. Fittanjem tog modela na podatke, dobivamo procjene $\bar{\theta}$ i tada stavljamo $\bar{F} = F(\bar{\theta})$.
- Primjer: kor.koef. za LAW school podatke
  1. Procjenimo kor. koef. $\bar{r}$ na stvarnim podacima
  2. Uz pretpostavku normalne distribucije za LSAT i GPA generiramo (LSAT, GPA) po bivarijatnoj normalnoj distribuciji sa korelacijskim koeficijentom $\bar{r}$.
  3. Na svakom generiranom uzorku procjenimo $\bar{r}^\star(b)$.
  4. Ponovimo korake 2-3 B puta.
  5. Izračunamo standardnu devijaciju $\bar{r}^\star(b)$ vrijednosti (na osnovu svih $b = 1 \ldots B$ vrijednosti)

Zadaća

8. zadaća: rok za predaju 3.6.
Zadaća se nalazi u folderu Zadaće na MERLINU.
UPUTE: Svaki zadatak iz zadaće mora biti u svom .sas programu. Sve .sas programe nazovite na način *prezime_ime_zad*1.*sas*, ako je npr. 1.zadatak u pitanju, itd. Sve što radite u zadaćama mora biti u obliku koda (možete koristiti sve dostupne materijale da dobijete tražene rezultate, ali sve mora biti napisano u obliku koda).