# PRIMER REŠEVANJA STATISTIČNIH PROBLEMOV Z BOOTSTRAP PRISTOPOM

Katarina Košmelj, Michael Schemper*

Univerza v Ljubljani, Slovenija

*Univerza na Dunaju, Avstrija

Predstavljene so osnove bootstrap tehnike, ki je zelo uporabna pri reševanju statističnih problemov, kjer parametrični pristopi niso možni.

Prikazan je problem izračunavanja intervala zaupanja za optimum kvadratne odzivne krivulje. Interval zaupanja je izračunan na parametrični in na neparametrični način. Parametrični način je v tem primeru le aproksimativen in temelji na razvoju v Taylorjevo vrsto, neparametrični način pa uporablja bootstrap tehniko.

Podana je ilustracija tega problema na podatkih iz agronomije. Rezultati kažejo, da se rezultati obeh pristopov precej razlikujejo in zdi se, da so rezultati dobljeni z bootstrap tehniko za uporabljene podatke bolj ustrezni.

KLJUČNE BESEDE: bootstrap tehnika, interval zaupanja za optimum odzivne krivulje.

AN APPLICATION OF THE BOOTSTRAP TECHNIQUE: The ideas of the bootstrap technique are briefly presented. An application is given when the confidence interval for the extreme of a quadratic response curve is calculated using the bootstrap technique. An alternative approach based on the asymptotic parametric approach (Taylor series) is presented as well. Both approaches are illustated on a data set from agriculture. It seems that the bootstrap technique gives more reliable results for our dataset.

KEYWORDS: bootstrap technique, confidence interval for the extreme of the response curve.

## 1. Uvod

Statistična inferenca je sklepanje o populaciji na osnovi vzorca. Da bi bil ta postopek korekten, moramo imeti reprezentativni vzorec, na osnovi katerega izračunamo oceno parametra, ki nas v populaciji zanima.

Kvaliteto vzorčne ocene izraža njena standardna napaka. Ta meri natančnost (angl. precision) vzorčne ocene in pove, kakšno napako naredimo, ko namesto celotne populacije obravnavamo le njen del (vzorec). (Pojem natančnosti se loči od pojma točnosti (angl. accuracy), ki upošteva poleg natančnosti tudi pristranskost).

Natančnost vzorčne ocene, torej njena standardna napaka, je pri statističnih analizah ključnega pomena. Obstojata dva načina izračunavanja standardne napake ocene:

- parametrični pristop, ki je možen, če poznamo ali predpostavljamo porazdelitev cenilke in znamo izračunati varianco ocene na osnovi te porazdelitve (klasični pristop statistične inference);

- neparametrični pristop, kadar porazdelitve ne poznamo in tako tudi variance ocene analitično ne moremo izračunati. Obstoja pa lahko asimptotično pravilna ocena.

## 2. Neparametrični pristop

Iz literature sta poznana dva neparametrična pristopa: bootstrap in jackknife pristop.

Opišimo na kratko bootstrap pristop. Predpostavimo, da smo s klasičnim vzorčenjem dobili reprezentativni vzorec, ki ima N enot. Imenujmo ga osnovni vzorec. Iz tega vzorca tvorimo t. im. bootstrap vzorce takole: vsak bootstrap vzorec ima N enot, dobimo pa ga s slučajnim izborom z vračanjem iz osnovnega vzorca. Tvorimo K bootstrap vzorcev. Iz vsakega bootstrap vzorca izračunamo vzorčno oceno. Na osnovi vseh K bootstrap vzorcev izdelamo porazdelitev vzorčne ocene. Če je porazdelitev vzorčne ocene približno normalna, izračunamo bootstrap oceno za povprečje in za varianco vzorčne ocene (glej formule spodaj). Navadno porazdelitev ni normalna, zato predhodno uporabimo primerno transformacijo.

Jackknife tehnika je v razliko od bootstrap tehnike deterministična. Posamezni jackknife vzorec dobimo tako, da spustimo po eno točko v osnovnem vzorcu. V vsakem jackknife vzorcu je N-1 točk, število jackknife vzorcev je torej enako N. To tehniko lahko interpretiramo kot aproksimacijo bootstrap tehnike.

Obe tehniki sta metodi ponovnega vzorčenja (angl. resampling) in predpostavljata, da je porazdelitvena funkcija F za oceno parametra, dovolj dobro ocenjena z empirično porazdelitveno funkcijo, ki jo dobimo na osnovi bootstrap vzorcev.

Naj bodo $x_1, x_2, \ldots x_N$ podatki. Poglejmo, kako izračunamo boostrap oceno za povprečje in za varianco vzorčne ocene. Naj bo T cenilka našega parametra. Iz i-tega bootstrap vzorca izračunamo $T_B^i = T(x_1^{(i)}, x_2^{(i)}, \ldots x_N^{(i)})$. Iz vseh K bootstrap vzorcev ocenimo vzorčno porazdelitev za T. Če je ta porazdelitev približno normalna, je ocena za povprečje $E_B(T)$ in za varianco $Var_B(T)$ po bootstrap tehniki

$$E_B(T) = \Sigma T_B^i/K$$

$$Var_B(T) = E_B(T^2) - E_B(T)^*E_B(T),$$

po jackknife tehniki pa

$E_J(T) = \Sigma T_j^J/N$

$Var_J(T) = (N-1)*(E_J(T^2) - E_J(T)*E_J(T))$

pri čemer je N število enot v osnovnem vzorcu, K pa število bootstrap vzorcev. Kot smo že omenili, v primeru, da porazdelitev ni 'dovolj normalna', uporabimo predhodno primerno transformacijo.

Empirično je bilo ugotovljeno, da naj bi bilo število bootstrap vzorcev K v primeru, ko želimo oceniti varianco vzorčne ocene, od 50 do 70. Bootstrap pristop se bolj obnese za majhne osnovne vzorce kot pa jackknife pristop. Za vrstilne karakteristike (npr. kvartili) je potrebno več ponovitev (K > 200), jackknife tehnika se izkaže v tem primeru zelo slabo.

Bootstrap in jackknife pristop sta izvedljiva le z ustrezno računalniško opremo. Sta algoritmično enostavna, vendar z vidika računalnišega časa zahtevna. Slednje pa postaja ob razvoju računalniške opreme vse manj pomembno.

## 3. Opis problema

Problem, ki ga predstavljamo, je izračun intervala zaupanja za optimum kvadratne odzivne krivulje. Zvezo med odvisno spremenljivko Y in pojasnjevalno spremenljivko X zapišemo

$E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2$

Na osnovi vzorca, v katerem je N točk, ocenimo parametre in izračunamo napovedi za odvisno spremenljivko

$\hat{Y} = b_0 + b_1 X + b_2 X^2$

Z odvajanjem dobimo oceno za lokacijo ekstrema

$x_{opt} = -b_1/2b_2$

Radi bi izračunali interval zaupanja za $x_{opt}$, za kar potrebujemo njegovo varianco

$Var(x_{opt}) = Var(-b_1/2b_2)$

Oceno za varianco $b = (b_0, b_1, b_2)$ dobimo iz variančno-kovariančne matrike

$Var(b) = (X'X)^{-1}\sigma^2$

Varianco za razmerje $-b_1/2b_2$ pa ne poznamo. Tako klasični parametrični pristop v tem primeru odpade.

### 3.1 Parametrični pristop

Razmerje razvijemo v Taylorjevo vrsto okoli točke $x_{opt}$ in poiščemo aproksimativno oceno za varianco. Upoštevamo le prve odvode (glej npr. Mead str. 532)

$$Var(x_{opt}) \doteq (1/4b_2^4)\,(b_2^2\,Var(b_1) + b_1^2\,Var(b_2) - 2b_1b_2\,Cov(b_1,b_2))$$

Opomnimo naj, da ne znamo ovrednotiti kvalitete te aproksimativne ocene.

### 3.2 Neparametrični pristop

Alternativni pristop k reševanju takih problemov so tehnike ponovnega vzorčenja.

## 4. Primer

### 4.1 Opis problema

Na Katedri za poljedeljstvo Biotehniške fakultete v Ljubljani so v letih 1989 in 1990 izvedli poljski poskus, s katerim so poskušali ugotoviti, pri kateri gostoti setve je pridelek koruze optimalen.

S tremi tipi setev so dosegli razmeroma široko območje gostote setve, kar je omogočilo dobro identifikacijo ekstrema odzivne krivulje. Poskus je bil zastavljen kot bločni poskus s 3 bloki, v vsakem bloku je bilo 15 različnih gostot. Skupaj imamo torej 45 meritev.

Grafični prikaz podatkov je na Sliki 1. V obeh primerih se je izkazalo, da se kvadratna krivulja dovolj dobro prilagaja podatkom, kar se je ujemalo s pričakovanji agronomov. Pri večjih gostotah (> 120) so se posebno v 1989 pokazale 'nestabilne' razmere, v obeh letih so se točke z največjo gostoto izkazale kot vplivne točke (angl. inference points). V Tabeli 1 so predstavljeni rezultati statistične analize.

Tabela 1: Predstavitev rezultatov prilagajanja kvadratne funkcije: ocene parametrov in njihove standardne napake, koeficient determinacije $R^2$, s, Durbin-Watsonova statistika DW, za posamezni leti.

| | 1989 | | 1990 | |
|---|---|---|---|---|
| | ocena | st. napaka | ocena | st. napaka |
| $b_0$ | 2077.73 | 559.69 | 249.19 | 321.88 |
| $b_1$ | 133.202 | 14.504 | 103.588 | 8.34109 |
| $b_2$ | -0.4667 | 0.0802 | -0.3854 | 0.04613 |
| $R^2$ | 70% | | 88% | |
| s | 1173.69 | | 1088.43 | |
| DW | 2.44 | | 1.95 | |

Izračunali smo oceni za maksimum pridelka iz odvoda. V poljedeljstvu točkovna ocena ne pove dosti, zato je izračun intervala zaupanja za maksimum s stališča stroke zaželen. Rezultati so v tabeli 2.

Tabela 2: Ocena za maksimum, za standardno napako (izračunano po aproksimativni analitični formuli) in 95% interval zaupanja za maksimum.

| | 1989 | 1990 |
|---|---|---|
| $x_{opt}$ | 121.3 | 134.4 |
| $s(x_{opt})$ | 7.30 | 6.48 |
| 95% IZ | (106.7, 135.9) | (127.9, 140.9) |

Intervala zaupanja se prekrivata na sorazmerno ozkem območju od 128 do 136.

Do ocene za standardno napako smo poskušali priti tudi z metodami ponovnega vzorčenja. Ker je bilo v osnovnem vzorcu 45 točk, smo se odločili za bootstrap pristop. Izdelali smo tri variante glede na število bootstrap vzorcev: K = 50, K = 100, K = 200.

Za vsak K (K = 50, 100 oz. 200) smo izvedli naslednji postopek: iz vsakega posameznega bootstrap vzorca izračunali optimum z odvajanjem, iz vseh K bootstrap vzorcev pa izdelali porazdelitev optimumov. Če je bila porazdelitev optimumov (vsaj) simetrična, smo izračunali povprečje in standardno napako za optimum. Če je bila porazdelitev asimetrična (to se je pogosto zgodilo za leto 1989), se s transformacijami nismo ukvarjali, ampak smo rezultate zavrgli in ponovili postopek vzorčenja.

Izdelali smo po 5 ponovitev za vsak K in za vsako leto. (Torej je bilo pri K = 200 za vsako leto 1000 ponovnih vzorčenj). Izračuni so bili izvedeni na zmogljivem velikem IBM računalniku in zanje smo pri K = 200 za vsako ponovitev porabili okoli 10 minut, pri K = 100 pa za vsako ponovitev okoli 5 minut. Rezultati so v Tabeli 3 in 4.

Tabela 3: povprečje in standardna napaka za optimum za 1989 glede na različno število bootstrap vzorcev K. Izdelanih je bilo po 5 ponovitev.

| 1989 | | | | | |
|---|---|---|---|---|---|
| K = 200 | | K = 100 | | K = 50 | |
| povprečje | st. napaka | povprečje | st. napaka | povprečje | st. napaka |
| 122.64 | 14.53 | 121.92 | 12.36 | 120.89 | 12.67 |
| 120.69 | 11.69 | 124.28 | 13.11 | 121.72 | 11.61 |
| 121.32 | 11.40 | 118.45 | 11.97 | 119.49 | 9.55 |
| 121.13 | 11.59 | 121.67 | 12.17 | 122.82 | 11.00 |
| 119.94 | 11.81 | 122.83 | 9.07 | 122.11 | 11.22 |

Tabela 4: povprečje in standarnda napaka za optimum za leto 1990 glede na različno število bootstrap vzorcev K. Izdelanih je bilo po 5 ponovitev.

| 1990 | | | | | |
|---|---|---|---|---|---|
| K = 200 | | K = 100 | | K = 50 | |
| povprečje | st. napaka | povprečje | st. napaka | povprečje | st. napaka |
| 134.20 | 5.40 | 134.55 | 5.41 | 133.93 | 3.87 |
| 134.28 | 4.92 | 133.82 | 4.77 | 133.36 | 6.20 |
| 133.93 | 4.96 | 134.24 | 5.07 | 133.19 | 4.13 |
| 134.91 | 4.83 | 133.58 | 4.06 | 133.97 | 4.59 |
| 134.49 | 4.73 | 133.88 | 5.30 | 135.37 | 4.62 |

Iz tabele 3 razberemo, da je za leto 1989 optimum približno 121, njegova standardna napaka pa približno 11. Za leto 1990 pa tabela 4 pokaže, da je optimum okoli 134, standardna napaka pa približno 5.

Primerjava: povprečji sta po obeh pristopih skoraj enaki (manj kot 1 % različni), različni pa sta standardni napaki: za 1989 daje aproksimativna ocena precej nižjo vrednost, za 1990 pa višjo vrednost. Zdi se, da so rezultati dobljeni po bootstrap metodi bolj kvalitetni, saj z večjo mero upoštevajo informacijo v podatkih.

Literatura:

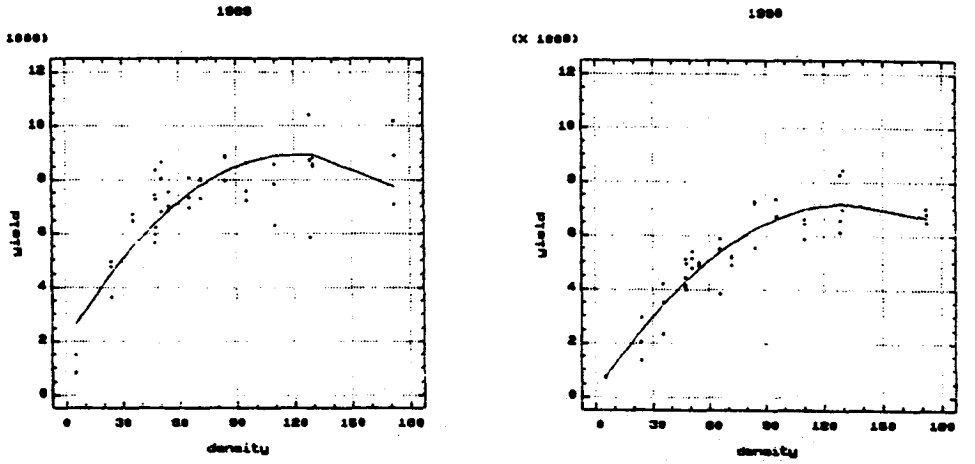

Bissel A.F.: The Jackknife - Toy, Tool or Two-edged Weapon?, The Statistician, Vol. 24, No. 2, 1975, p. 79-100

Efron B.: Nonparametric standard errors and confidence intervals, The Canadian Journal of Statistics, Vol. 9, No. 2, 1981, p. 139-172

Mead R.: The Design of Experiments. Statistical principles for practical applications, Cambridge University Press, 1988

Schemper M.: Some Applications of Bootstrap Techniques to Nonparametrics, neobjavljeno

Weihs Claus: Canonical Discriminant Analysis: Comparison of Simulation Methods, Seminar 91 der Region Oesterreich-Schweiz der Internationalen Biometrischen Geselschaft, Biel, 1991




Slika 1: Grafični prikaz podatkov in odzivne krivulje za leto 1989 in 1990: gostota v 1000 rastlin/ha, pridelek v tonah