# Bootstrap metoda u prilagodbi linearnog regresijskog modela

**Tomić, Bernarda**

**SVEUČILIŠTE U ZAGREBU**

**PRIRODOSLOVNO–MATEMATIČKI FAKULTET**

**MATEMATIČKI ODSJEK**

Bernarda Tomić

# BOOTSTRAP METODA U PRILAGODBI LINEARNOG REGRESIJSKOG MODELA

Diplomski rad

Voditelj rada:
Prof.dr.sc. Miljenko Huzak

Zagreb, veljača, 2015. godina

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

*Hvala svima koji su imali strpljenja i tolerancije za mene, posebice mentoru i obitelji.*
*Također i onima koji će imati.*

# Sadržaj

# Predgovor

U svom diplomskom radu bavit ću se bootstrap metodom za procjenu parametara uzoračkih razdioba određenih statistika (na primjer, pouzdanih intervala za parametre modela, standardne pogreške procjenitelja i slično). Objasnit ću osnovne postavke i principe metode te njenu primjenu u analizi prilagodbe linearnih regresijskih modela u situacijama kada ne vrijede uobičajene pretpostavke o pogreškama (normalnost, homoskedastičnost, nekoreliranost i sl.).

Bradley Efron, otac bootstrapa, odabrao je izraz 'bootstrap'za svoju shemu ponovljenog uzorkovanja (resampling shemu) motiviran sljedećim: termin proizlazi iz fraze *to pull oneself up by one's own bootstraps = vlastitim silama*, što pak dolazi iz knjige *The surprising adventures of Baron Munchausen (1785.)*, citat str 22. :

*I was still a couple of miles above the clouds when it broke, and with such violence I fell to the ground that I found myself stunned, and in a hole nine fathoms under the grass, when I recovered, hardly knowing how to get out again. Looking down, I observed that I had on a pair of boots with exceptionally sturdy straps. Grasping them firmly, I pulled with all my might. Soon I had hoist myself to the top and stepped out on terra firma without further ado.*

Osnovna ideja metode je generirati uzorak iz već postojećeg i na temelju generiranih uzoraka procijeniti određene statističke veličine. Budući da se metoda pokazala uspješnom u rješavanju mnogih statističkih problema prihvaćena je kao alternativa asimptotskim metodama (ponekad i bolja od nekih, na primjer standardne normalne aproksimacije i Edgeworthove ekspanzije).

U idućoj točki reći ću ponešto o metodama ponovljenog uzorkovanja (resampling metodama), zatim slijede opis ideje bootstrap metode i jednostavan primjer kako bismo vidjeli njenu primjenu. U drugom poglavlju bit će više riječi o pojedinim vrstama bootstrapa i opravdanosti korištenja metode. Treće poglavlje sadrži opis bootstrapa u linearnom regresijskom modelu, a na kraju ću napomenuti kada primijeniti metodu.

# Poglavlje 1

# Uvod

## 1.1 Metode ponovljenog uzorkovanja

Metode ponovljenog uzorkovanja sastavni su dio analize podataka. Njihova prednost u usporedbi sa standardnim metodama statističkog zaključivanja su veće jednostavnost i točnost, zahtjev za manje pretpostavki i veća općenitost. Očita prednost je kada nisu zadovoljene osnovne pretpostavke tradicionalnih parametarski testova, što je slučaj s malim uzorcima i nenormalnim distribucijama. Dodatno, navedene metode mogu dati odgovore na pitanja kao što su usporedba medijana ili proporcija, za razliku od tradicionalnih metoda. One su iste za testiranje sredina, medijana, proporcija i ostalih parametara što ih čini konceptualno zaista jednostavnima. Glavne vrste metoda su: randomizacijski test, kros-validacija, jackknife i bootstrap.

Povijesno, sve je počelo s jackknife metodom pa ću od nje i krenuti, na primjeru gdje zapravo i nije nužno njeno korištenje. Procijenimo standardnu devijaciju uzoračke sredine. Skup podataka (zapravo slučajni uzorak) sastoji se od $n$ nezavisnih i jednako distribuiranih slučajnih varijabli nepoznate distribucije $F$,

$$X_1, X_2, ..., X_n \sim F. \tag{1.1}$$

Iz jedne realizacije slučajnog uzorka, $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ možemo izračunati uzoračku sredinu $\bar{x} = \sum_{i=1}^{n} x_i/n$ kao nepristranu procjenu očekivanja od $F$. Zanimljiva činjenica, krucijalna za statističku primjenu, jest da uzorak osim procjene za očekivanje daje i procjenu $\widehat{\sigma}$ za njenu točnost:

$$\widehat{\sigma} = \Big[\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2\Big]^{1/2}, \tag{1.2}$$

gdje je $\widehat{\sigma}$ procijenjena standardna devijacija od $\bar{X}$. Problem s formulom (1.2) je taj da se na niti jedan očiti način ne može proširiti na druge procjenitelje osim na $\bar{X}$. Koristeći jacknife možemo na primjer dobiti i procjenu uzoračkog medijana. Neka je

$$\bar{x}_{(i)} = \frac{n\bar{x} - x_i}{n-1} = \frac{1}{n-1} \sum_{j \neq i} x_i, \tag{1.3}$$

uzoračka sredina danog skupa podataka bez $i$-tog elementa. Također, neka je $\bar{x}_{(.)} = \sum_{i=1}^{n} x_{(i)}/n$ prosjek izbrisanih sredina. Jackknife procjena standardne devijacije jest

$$\widehat{\sigma}_{JACK} = \Big(\frac{n-1}{n} \sum_{i=1}^{n} (\bar{x}_{(i)} - \bar{x}_{(.)})^2\Big)^{1/2}. \tag{1.4}$$

Ovo je u biti isto kao i (1.2). Prednost jednadžbe (1.4) je što se može poopćiti na bilo koji procjenitelj $\widehat{\theta} = \widehat{\theta}(X_1, X_2, ..., X_n)$. Jedina promjena je u zamjeni $\widehat{\theta}_i = \widehat{\theta}(X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n)$ za $\bar{x}_{(i)}$ i $\widehat{\theta}_{(\cdot)}$ za $\bar{x}_{(\cdot)}$.

Bootstrap generalizira (1.3) na drugačiji način. Neka je $\widehat{F}$ empirijska funkcija distribucije podataka (pridružuje $1/n$ svakom $x_i$-u) i neka su $X_1^*, X_2^*, ..., X_n^*$ nezavisne, jednako distribuirane slučajne varijable iz $\widehat{F}$,

$$X_1^*, X_2^*, ..., X_n^* \sim \widehat{F}, \tag{1.5}$$

odnosno $X_i^*$ je slučajni uzorak (izvučen s mogućim ponavljanjima iz promatranih vrijednosti $x_1, ..., x_n$). Tada $\bar{X}^* = \sum(X_i^*/n)$ ima varijancu

$$Var_* \bar{X}^* = \frac{1}{n^2} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{1.6}$$

gdje $Var_*$ označava varijancu bootstrap uzorka (1.5). Bootstrap procjena standardne devijacije za procjenitelj $\widehat{\theta}(X_1, X_2, ..., X_n)$ je

$$\widehat{\sigma}_{BOOT} = \Big[Var_* \widehat{\theta}(X_1^*, X_2^*, ..., X_n^*)\Big]^{1/2}. \tag{1.7}$$

Usporedbom (1.7) i (1.2) vidimo da je $\widehat{\sigma}_{BOOT} = [(n-1)/n]^{1/2} \widehat{\sigma}$ za $\widehat{\theta} = \bar{X}$. Moguće je dobiti i da je $\widehat{\sigma}_{BOOT} = \widehat{\sigma}$ za $\widehat{\theta} = \bar{X}$ jednostavnim dodavanjem faktora $[n/(n-1)]^{1/2}$ u definiciju (1.7).

Sada ću reći ponešto o kros-validaciji i randomizacijskom testu. Randomizacijski test je razvio R.A. Fisher (1935.), utemeljitelj klasičnog statističkog testiranja. U testu se na slučajan način realociraju podaci te se za takve permutirane podatke računa p - vrijednost. Na primjer, uspoređujemo sljedeće rezultate nekog eksperimenta : $\{99, 90, 93, ...\}$ sa $\{87, 89, 97, ...\}$. Neka je rezultat t-testa t-vrijednost 1.55. Klasičnom metodom određivali bismo je li razlika grupa značajna usporedbom $t_{uzorak}$ i $t_{teorijski}$ u t - distribuciji. U metodi ponovljenog uzorkovanja, umjesto usporedbe s teoretskom distribucijom, mijenjajući realizacije između grupa, mijenja se i t - vrijednost. Svrha ovog postupka je simulacija "šanse". Nakon ponavljanja postupka na primjer 100 puta, ako se t-vrijednost 1.55 pojavljuje u 5% slučajeva možemo zaključiti da je egzaktna p-vrijednost 0.05.

U kros-validaciji uzorak se dijeli na dva ili više poduzorka i onda se model testira na svakom od njih. Cilj je zaključiti je li pojedini rezultat ponovljiv ili je samo stvar slučajnih fluktuacija. Uzmimo regresiju za primjer. U kros-validaciji se prvi poduzorak obično uzima za izvod jednadžbe regresije, idući za generiranje prediktivnih rezultata iz nje. Kros-validacijski koeficijent se računa korelacijom prediktivnih rezultata i promatranih grešaka rezultatnih varijabli. Očito ovdje nema ponovnog korištenja istih podataka.

Dakle, kros-validacija izvorno služi provjeri ponovljivosti rezultata, jackknife omogućuje otkrivanje outliera, a bootstrap donošenje zaključaka. Poklonici metoda navode nekoliko razloga kojima opravdavaju spomenute tehnike:

- Empirijski. Klasični postupci se oslanjaju na teoretske distribucije koje zahtjevaju jake pretpostavke o uzorku i populaciji. Kod krivih pretpostavki o distribuciji dolazimo do grešaka u zaključivanju. U slučaju kada smo skeptični u vezi upotrebe teoretske distribucije, empirijski bazirana metoda ponovljenog uzorkovanja je dobra alternativa.

- Jasnoća. Konceptualno govoreći, ponovljeno uzorkovanje je čisto i jednostavno. Sofisticirana matematička pozadina nije nužna za njegovo shvaćanje što omogućuje bavljenje sadržajem istraživanja umjesto vođenja brige o tome koji test uzeti.

- Distribucija. Klasične metode zahtjevaju određene pretpostavke o distribuciji, međuostalima i o veličini uzorka. Kada je on nedovoljno velik, rješenje je metoda ponovljenog uzorkovanja.

- Veliki uzorak. Obično se ponovljeno uzorkovanje koristi kada je uzorak mali, međutim može se upotrijebiti i u ovom slučaju. Ako imamo jako veliki uzorak, moguće je odbiti bilo koju nul hipotezu. U tom slučaju moguće je podijeliti uzorak na nekoliko podskupova i primijeniti kros-validaciju.

- Ponavljanja. Klasične metode ne daju odgovor koliko je vjerojatno da se rezultati ponove. Kros-validacija i bootstrap mogu se koristiti u tu svrhu.

## 1.2 Glavna ideja bootstrap metode

Kao što sam već spomenula u predgovoru, osnovna ideja metode je ponovljeno uzorkovanje nezavisnih događaja (resampling) kako bismo procijenili određene statističke parametre. Ovakav pristup zapravo uključuje ponavljanje izvorne analize sa puno skupova podataka generiranih iz početnog skupa. Metoda je primjenjiva neovisno o tome imamo li definiran vjerojatnosni model za podatke. Korisna je u slučajevima kada bi nam se zaključak bazirao na kompleksnim postupcima za koje su teoretski rezultati nedostupni ili neprimjenjivi zbog nedovoljno velikog uzorka, a kada je standardni model pretpostavljen, ali nije jasno s čime ga zamijeniti ili pak kada tražimo 'quick and dirty'odgovor. Metoda može biti i jako korisna za provjeru korisnosti standardne aproksimacije za parametarske modele ili pak za poboljšanje ako model daje neadekvatne zaključke. Ako na primjer ne znamo koji je prikladni procjenitelj za standardnu grešku populacijskog parametra ili pak sumnjamo da bi naš procjenitelj mogao biti pristran, podaci su nam rijetki i pretpostavke nesigurne, možemo koristiti alate poput bootstrapa. Kako bih slikovitije opisala metodu, navest ću i nekoliko primjera realiziranih pomoću R-a, programskog jezika i statističkog softvera za analizu podataka, modeliranje i grafiku.

Najjednostavniji primjer je kada proučavane podatke (uzorak) $y_1, ..., y_n$ tretiramo kao realizaciju slučajnog uzorka $Y_1, ..., Y_n$ iz nepoznate razdiobe $F$. Fokus je stavljen na parametar $\theta$ koji je realizacija statističkog funkcionala $t(\cdot)$ na $F$, odnosno $\theta = t(F)$. Trivijalni primjer takvog funkcionala je sredina, $t(F) = \int ydF(y)$. Općenito o $t(\cdot)$ mislimo kao o nekom algoritmu koji namjeravamo primijeniti na $F$.

Procjenitelj od $\theta$ je $\widehat{t} = t(\widehat{F})$, gdje je $\widehat{F}$ procjenitelj od $F$ (na temelju uzorka $y_1, ..., y_n$). To može biti parametarski model, kao što je na primjer normalni, s parametrima procijenjenim metodom maksimalne vjerodostojnosti ili nekom robusnijom metodom, ili pak empirijska funkcija distribucije (EFD) $\widehat{F}$, koja pridaje masu $n^{-1}$ svakom $y_i$.
Kako god dobili $\widehat{F}$ naš procjenitelj $\widehat{t} = t(\widehat{F})$ je jednostavno realizacija funkcionala $t$ na $\widehat{F}$.

Tipični problemi koji se pojavljuju su: kolike su pristranost (*bias*) i procjena varijance za $\widehat{t}$? Koji je vjerodostojni pouzdani interval za $\theta$? Je li određena hipoteza konzistentna s podacima?

Ja ću se uglavnom fokusirati na pouzdane intervale. Najjednostavniji pristup kod konstrukcije pouzdanih intervala jest korištenje normalne aproksimacije kao distribucije od $T$, slučajne varijable čija je $\widehat{t}$ jedna realizacija. Ako su stvarne vrijednosti pristranosti i varijance po definiciji redom:

$$b(F) = E(T|F) - \theta = E(T|F) - t(F), v(F) = var(T|F), \tag{1.8}$$

onda je za velike uzorke

$$Z = \frac{T - t(F) - b(F)}{v(F)^{1/2}} \dot{\sim} N(0,1),$$

gdje $\dot{\sim}$ znači "asimptotski". Uvjet pod (1.1) indicira da je $T$ bazirana na slučajnom uzorku $Y_1, ..., Y_n$ iz $F$. U ovom slučaju aproksimativni $(1-2\alpha)\%$ pouzdani interval za $\theta = t(F)$ jest

$$[\widehat{t} - b(F) - z_{1-\alpha} v(F)^{1/2}, \widehat{t} - b(F) - z_{\alpha} v(F)^{1/2}], \tag{1.9}$$

gdje je $z_\alpha$ $\alpha$ kvantil standardne normalne razdiobe. Kako je $F$ nepoznat, ideja bootstrapa je da ga zamijenimo s poznatim procjeniteljem $\widehat{F}$ te procijenimo $b(\widehat{F})$ i $v(\widehat{F})$, što je moguće analitički samo u najjednostavnijim slučajevima pa zato koristimo simulaciju. Generiramo $R$ nezavisnih bootstrap uzoraka $Y_1^*, ..., Y_n^*$ nezavisnim uzorkovanjem iz $\widehat{F}$, izračunamo odgovarajući procjenitelj slučajnih varijabli $T_1^*, ..., T_n^*$ te se nadamo da približno vrijedi

$$b(F) \doteq b(\widehat{F}) = E(T|\widehat{F}) - t(\widehat{F}), \tag{1.10}$$

$$b(F) \doteq R^{-1} \sum_{r=1}^{R} T_r^* - \widehat{t} = \overline{T}^* - \widehat{t}, \tag{1.11}$$

$$v(F) \doteq v(\widehat{F}) = var(T|\widehat{F}), \tag{1.12}$$

$$v(F) \doteq \frac{1}{R-1} \sum_{r=1}^{R} (T_r^* - \overline{T}^*)^2, \tag{1.13}$$

gdje $\doteq$ označava "približno jednako". Ovdje su prisutne dvije greške: statistička uslijed zamjene $F$ sa $\widehat{F}$ i pogreška simulacije zbog zamjene očekivanja i varijance uzoračkim sredinama. Ako je moguće, koristimo $b(\widehat{F})$ i $v(\widehat{F})$ na takav način da statističku pogrešku, koja je neizbježna u većini situacija, minimiziramo.

Ako ne možemo koristiti pretpostavke o normalnoj distribuciji, alternativni pristup određivanja pouzdanih intervala bi se mogao bazirati na $T - \theta$. Ideja je sljedeća: ako $T^* - \widehat{t}$ i $T - \theta$ imaju približno istu distribuciju, onda kvantili druge mogu biti procijenjeni simulacijom kvantila iz prve distribucije, čime dobivamo osnovni bootstrap $(1-2\alpha)\%$ pouzdani interval

$$[\widehat{t} - (T^*_{(R+1)(1-\alpha)} - \widehat{t}), \widehat{t} - (T^*_{(R+1)\alpha} - \widehat{t})],$$

gdje su $T_1^*, ..., T_R^*$ uređajne statistike od $T_r^*$.Kada iz $Y_1, ..., Y_n$ možemo izračunati aproksimativnu varijancu $V$ za $T$, studentizirani bootstrap pouzdani interval se temelji na $Z = (T-\theta)/V^{1/2}$, čiji kvantili su procijenjeni iz simuliranih vrijednosti uređajne statistike bootstrap uzorka $T_1^*, ..., T_R^*$ odgovarajuće bootstrap varijable $Z^* = (T^* - \widehat{t})/V^{*1/2}$. Iz empirijske funkcije distribucije od $Z_1^*, ..., Z_R^*$ računamo kvantile $Z^*_{(R+1)(1-\alpha)}$ i $Z^*_{(R+1)(\alpha)}$ . Studentizirani bootstrap interval tada je dan sa

$$[\widehat{t} - V^{1/2}Z^*_{(R+1)(1-\alpha)}, \widehat{t} - V^{1/2}Z^*_{(R+1)(\alpha)}].$$

Sada ću jednostavnim primjerom pokazati opisanu ideju koristeći neparametarski boostrap (kasnije će biti više riječi kako o njemu tako i o parametarskom i poluparametarskom bootstrapu).

## 1.3 Motivacijski primjer

Želimo dobiti procjenu intervala za $\gamma$ -modificiranu sredinu populacije iz koje imamo izvučenih 12 brojeva na slučajan način. U našem primjeru to je niz $n = (0, 1, 2, 3, 4, 6, 8, 10, 10, 12, 13, 15)$. $\gamma$-modificirana sredina je definirana sa:

$$\mu_t = \frac{1}{1-2\gamma}\int_{x_y}^{x_{1-\gamma}} xdF(x),$$

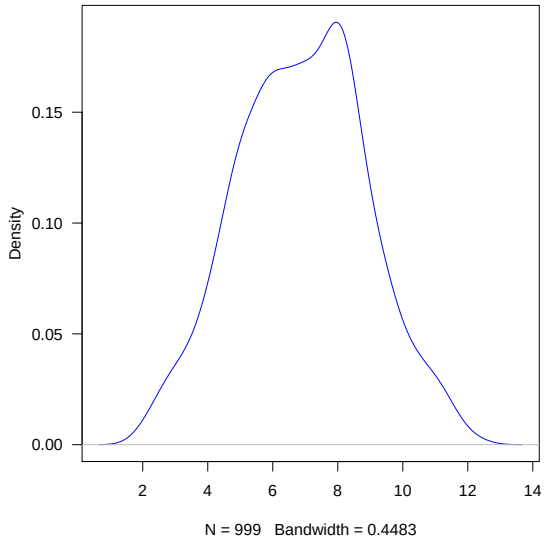
gdje je $X$ slučajna varijabla distribucije $F$. Riječima, $\mu_t$ je sredina distribucije nakon što je ona modificirana na sljedeći način: odrežemo $\gamma$ i $1-\gamma$ kvantil (mi ćemo pretpostavljati dvostrano rezanje iako se negdje pretpostavlja i jednostrano). Općenito, kada govorimo o modificiranoj distribuciji, želimo reći da se vjerojatnosna funkcija gustoće $f(x)$ transformira u

$$\frac{1}{1-2\gamma}f(x), x_\gamma \leq x \leq x_{1-\gamma},$$

gdje su $x_\gamma$ i $x_{1-\gamma}$ redom $\gamma$ i $1-\gamma$ kvantili. Primjer koji ću opisati je preuzet iz [7]. Za implementaciju koristimo library *boot*.

Procijenimo 25% ($\gamma = 0.25$) modificiranu sredinu populacije iz koje smo uzeli uzorak $n$. Koristimo naredbu *mean* s vrijednosti 0.25 parametra *trim*. Dobivena točkovna procjena je 6.8333. Sada ćemo smatrati da je uzorak populacija i provesti ponovljeno uzorkovanje nezavisnih događaja (s ponavljanjem) te nakon toga opet izračunati 25% ($\gamma = 0.25$) modificiranu sredinu. Kako se moglo i očekivati, rezultat je drugačiji (dobivena vrijednost je 4). Jedna iteracija *ponovno uzorkovanje pa procjena* nije od pomoći. Ako pak ponovimo opisani postupak na primjer 1000 puta, dobit ćemo distribuciju koja procjenjuje uzoračku distribuciju procjenitelja modificirane sredine populacije. (Slika 1.1)

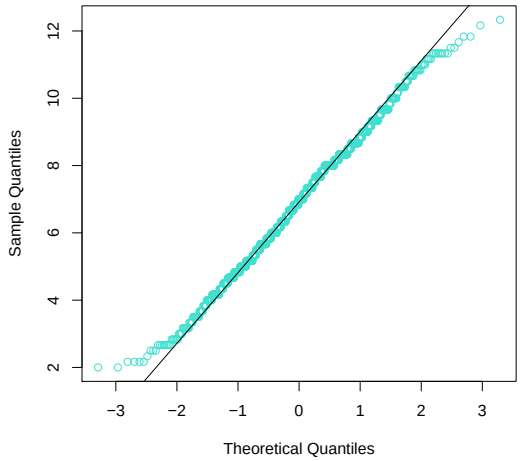
Distribucija na jednostavan način sugerira procjenu i korekciju eventualnih pristranosti u dobivenom procjenitelju razlikom između izvornog (modificirana sredina uzorka $n$) procjenitelja i sredine bootstrap uzorka.




Slika 1.1: Uzoračka distribucija modificirane sredine

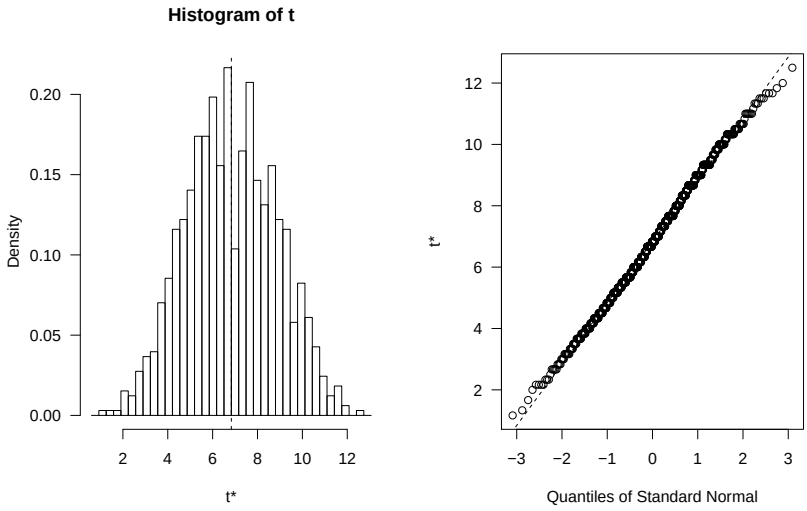
Ako želimo odrediti procjenu intervala na ovaj način, dodajemo izvornoj procjeni razliku između izvorne procjene i gornjeg percentila odgovarajućeg bootstrap uzorka, i obratno. Ovako dobiveni interval se zove osnovni bootstrap interval. Distribucija na Slici 1.1 također sugerira jednostavan način za dobivanje neparametarske intervalne procjene za modificiranu sredinu. Možemo jednostavno odrediti 2.5% i 97.5% percentile procijenjene gustoće. Takav pristup zovemo procjena percentila. Iako se intuitivno čini zadovoljavajuć, on ne ispravlja eventualnu pristranost. Ako prepostavimo da je naša uzoračka distribucija stvarno normalna, onda bi imalo smisla procijeniti standardnu pogrešku procjene i konstruirati pouzdane intervale, koji su bitno stabilniji od procjene percentila. Procjenjujemo standardnu pogrešku koristeći standardnu devijaciju uzoračke distribucije. Ovaj pristup konstrukcije intervala procjene zovemo normalna procjena. Možemo provjeriti je li normalnost opravdana koristeći poznat alat, normalni kvantilni graf za testiranje normalnosti.

Na Slici 1.2 vidimo da je pretpostavka o normalnosti uzoračke distribucije opravdana.



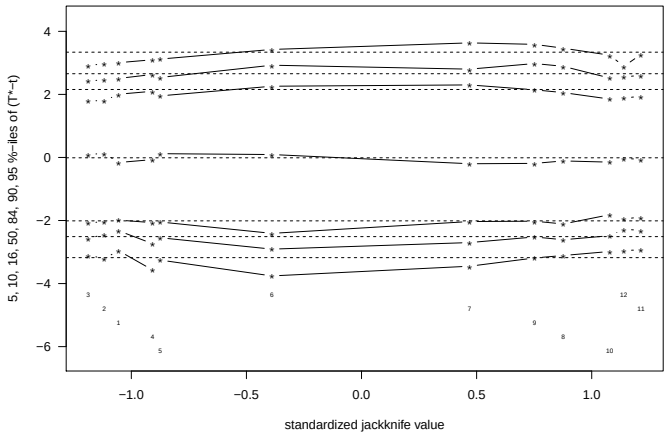

Slika 1.2: Graf kvantila normalne razdiobe procjenjene uzoračke distribucije 25% modificirane sredine

Za formalnu implementaciju bootstrapa treba nam funkcija najmanje dva parametra: uzorka i indeksa koji će ga permutirati. Funkcija vraća 25% modificiranu sredinu permutacije. Koristeći *boot* funkciju za podatke iz primjera i dobivenu modificiranu sredinu u 999 ponavljanja, dobijemo bootstrap output na Slici 1.3, koji opet opravdava pretpostavku o normalnosti. Bootstrap uzorak nam daje informacije o izvornoj procjeni te procjeni pristranosti i standardne greške procjenitelja. Pouzdane intervale osnovnom, percentilnom i normalnom procjenom dobijemo funkcijom *boot.ci*. Kao što sam i prije spomenula postoji još metoda za njihovo određivanje, na primjer studentizirani bootstrap (koristimo u slučaju kada možemo procijeniti varijancu procjenitelja za svaki bootstrap uzorak), bootstrap korigiran za pristranost (u slučaju kada je je procjenitelj pristran ili distribucija diskretna) itd.




Slika 1.3: Dijagnostički graf za 25% modificiranu bootstrap sredinu

Uz histogram i graf kvantila normalne razdiobe često nas zanima doprinos svake od opservacija na cjelokupnu sliku. Ovu informaciju možemo dobiti iz bootstrapa koristeći takozvani *jackknife-after-bootstrap* alat.




Slika 1.4: Jackknife-after-bootstrap grafički dijagnostički alat

Za svaku opservaciju razmatrani su bootstrap uzorci iz kojih je isključena. Percentili njihovih procjena su zatim grafički prikazani kao niz točaka s brojevima opservacije ispod, na x-lokaciji koja odgovara procjeni njihovog efekta na ukupnu bootstrap procjenu. Horizon-

talne točkaste linije predstavljaju ukupnu bootstrap distribuciju. Oštra devijacija od ovih linija sugerira točku koja ima znatan utjecaj na ishod. Za skup podataka iz ovog primjera nijedna točka ne iskače posebno. Za kraj uvoda, često pitanje je: koliko bootstrap ponavljanje bismo trebali koristiti? Općenito, ona su vrlo jeftina i tada je odgovor u tisućama. Ako su pak skuplja, tada kolikogod nam vremenska ograničenja omogućuju. Unatoč atraktivnim svojstvima, bootstrap nije uvijek efikasan. Posebno bismo trebali biti oprezni kada su podaci koje posjedujemo nepotpuni ili *dirty* ili kada je pretpostavka o njihovoj nezavisnoti upitna.
Bootstrap će nam dati odgovor u svim pa i tim slučajevima, ali tada bi on mogao biti nepouzdan.

# Poglavlje 2

# Bootstrap metoda

## 2.1 Parametarski i neparametarski bootstrap

Prvo ću ponoviti ukratko dio notacije i dodati novi. Uzorak $X = \{X_1, ..., X_n\}$ je kolekcija $n$ slučajno generiranih brojeva iz populacije, odnosno $X_i$ su nezavisne i jednako distribuirane slučajne varijable, od kojih svaka ima populacijsku funkciju distribucije $F$. U takozvanim neparametarskim problemima, ponovljeni uzorak $X^* = \{X_1^*, ..., X_n^*\}$ je neuređena kolekcija $n$ elementata slučajno generiranih iz $X$, s mogućim ponavljanjima. Svaki $X_i^*$ ima vjerojatnost $n^{-1}$ da bude jednak nekoj realizaciji $x_j$ iz $X$.

$$P(X_i^* = x_j|X) = n^{-1}, 1 \leq i, j \leq n.$$

Vidimo iz gornje jednadžbe da su varijable $X_i^*$ nezavisne i jednako distribuirane, uvjetno na $X$. $X^*$ će vjerojatno imati elemente koji se ponavljaju. U parametarskim problemima $X^*$ označava slučajno generirani uzorak iz populacije kojoj se procjenjuje parametar. Ako je ta populacija neprekidna, tada je vjerojatnost da su svi elementi iz $X^*$ različiti jednaka jedan. U oba slučaja (parametarskom i neparametarskom), $\widehat{F}$ označava distribucijsku funkciju "*populaciji*" iz koje je $X^*$ generiran. Par $(F, \widehat{F})$ koji označava populacijsku i uzoračku funkciju distribucije redom će često biti zapisan kao $(F_0, F_1)$, kako bi se lakše prikazala bootstrap iteracija, u kojoj za neki $i \geq 1$, $F_i$ označava funkciju distribucije uzorka generiranog iz $F_{i-1}$ uvjetno na $F_{i-1}$. Iz istog razloga, par $(X, X^*)$ će često biti pisan kao $(X_1, X_2)$.

Procjena $\widehat{\theta}$ je funkcija slučajnog uzorka i može se promatrati kao funkcional uzoračke distribucije funkcije $\widehat{F}$. Te dvije funkcije su numerički ekvivalentne, ali korisno ih je razlikovati, što možemo na sljedeći način: $\widehat{\theta} = \theta[X] = \theta(\widehat{F})$. Mnogi statistički problemi mogu se zapisati u obliku populacijske jednadžbe. Bootstrap procjena rješenja takve jednadžbe dobivena je rješavanjem uzoračke jednadžbe. Formalno, za dani funkcional $f_t$ iz klase $\{f_t : t \in T\}$ želimo odrediti vrijednost $t_0$ od $t$ koja je rješenje jednadžbe

$$E\{f_t(F_0, F_1)|F_0\} = 0, \tag{2.1}$$

gdje $F_0$ označava populacijsku, a $F_1$ uzoračku funkciju distribucije. Jednadžbu (2.1) zovemo populacijska jednadžba jer nam trebaju svojstva populacije da bismo ju točno riješili. Neka je sada $F_2$ funkcija distribucije uzorka iz $F_1$ (uvjetno na $F_1$). Sada jednostavno zamijenimo par $(F_0, F_1)$ sa $(F_1, F_2)$ transformirajući (2.1) u

$$E\{f_t(F_1, F_2)|F_1\} = 0. \tag{2.2}$$

Potonju zovemo uzoračka jednadžba jer znamo sve o njoj jednom kada znamo uzoračku distribuciju funkcije $F_1$. Posebno, rješenje jednadžbe $\widehat{t_0}$ je funkcija uzoračkih vrijednosti. Ideja je, naravno, da je rješenje uzoračke jednadžbe dobra aproksimacija rješenja populacijske jednadžbe. Ovo nazivamo "bootstrap principom".

Sada idemo na konstrukciju simetričnog 95% pouzdanog intervala za $\theta_0$, gdje je $\theta_0 = \theta(F_0)$ stvarna vrijednost parametra $\theta$. Rješavamo (2.1) kada je

$$f_t(F_0, F_1) = I\{\theta(F_1) - t \leq \theta(F_0) \leq \theta(F_1) + t\} - 0.95. \tag{2.3}$$

Indikator $I(\epsilon)$ je jednak 1 ako događaj $\epsilon$ ostvaren, 0 inače. Pouzdani interval je oblika $(\widehat{\theta} - t_0, \widehat{\theta} + t_0)$, gdje je $\widehat{\theta} = \theta(F_1)$. Kao što sam spomenula u uvodu, ovo zovemo (simetrična) percentilna metoda pouzdanog intervala za $\theta_0$. $\widehat{t_0}$ i $E\{f_t(F_1, F_2)|F_1\}$ su bootstrap procjenitelji za $t_0$ i $E\{f_t(F_0, F_1)|F_0\}$ redom.

Prikladno je dati detaljniju definiciju od $F_1$ i $F_2$. Osnovna dva pristupa su ona prikladna za parametarske, odnosno neparametarske modele. U oba slučaja zaljučivanje nam se bazira na uzorku $X$, koji sadrži $n$ slučajnih (nezavisnih i jednako distribuiranih) opservacija. U neparametarskom slučaju $F_1$ je jednostavno empirijska funkcija distribucije od $X$, odnosno funkcija koja svakoj točki u $X$ pridružuje masu $n^{-1}$. Slično, $F_2$ je empirijska funkcija distribucije uzorka iz populacije koja ima funkciju distribucije $F_1$. Taj uzorak označavamo s $X^*$. Ako označimo populaciju s $X_0$, tada imamo ugniježđeno uzorkovanje na način da je $X$ uzorak iz $X_0$, a $X^*$ iz $X$.

U parametarskom slučaju pretpostavljamo da je $F_0$ potpuno poznata i definirana konačnim vektorom $\lambda_0$ nepoznatih parametara. Ovisnost označavamo sljedećom jednadžbom $F_0 = F(\lambda_0)$, koja je element klase $\{F_\lambda, \lambda \in \Lambda\}$ mogućih distribucija. Neka je $\widehat{\lambda}$ procjena od $\lambda_0$ dobivena iz $X$, često metodom maksimalne vjerodostojnosti. To će biti funkcija uzoračkih vrijednosti pa možemo pisati $\lambda[X]$. Tada je $F_1 = F_{\widehat{(\lambda)}}$, funkcija distribucije koju dobijemo zamjenom stvarnih vrijednosti parametra s uzoračkim procjenama. Neka je $X^*$ uzorak iz $X$ sa funkcijom distribucije $F_{\widehat{(\lambda)}}$ i neka $\widehat{\lambda}^* = \lambda[X^*]$ označava verziju od $\widehat{\lambda}$ izračunatu iz $X^*$ umjesto iz $X$. Tada je $F_2 = F_{\widehat{(\lambda^*)}}$.

U oba slučaja, parametarskom i neparametarskom, $X^*$ je dobiven ponovljenim uzorkovanjem iz distribucije definirane originalnim uzorkom $X$. Sada ću primjerom detaljnije ilustrirati napisano.

Simetrični pouzdani interval za $\theta_0 = \theta(F_0)$ može se konstruirati primjenom metode ponovljenog uzorkovanja koristeći funkciju $f_t$ definiranu sa (2.3). Uzoračka jednadžba tada izgleda ovako:

$$P\{\theta(F_2) - t \leq \theta(F_1) \leq \theta(F_2) + t|F_1\} - 0.95 = 0. \tag{2.4}$$

U neparametarskom kontekstu, $\theta(F_2)$ uvjetno na $F_1$ ima diskretnu distribuciju pa bi rijetko bilo moguće točno riješiti (2.4). Ipak, eventualne pogreške su prilično male budući da se najveći atom distribucije $\theta(F_2)$ smanjuje eksponencijalno brzo kako $n$ raste. Mogli bismo i olakšati proces povećavanjem glatkoće funkcije distribucije $F_1$. O tome možete pročitati više u 4. poglavlju od [1]. U parametarskom slučaju, (2.4) se uglavnom može točno riješiti za $t$. S obzirom da su greške jako male i brzo se smanjuju kako $n$ raste, možemo reći da je

$$\widehat{t_0} = inf\{t : P[\theta(F_2) - t \leq Q(F_1) \leq Q(F_2) + t|F_1] - 0.95 \geq 0\} \tag{2.5}$$

rješenje od (2.4). $(\widehat{\theta} - \widehat{t_0}, \widehat{\theta} + \widehat{t_0})$ je bootstrap pouzdani interval za $\theta_0 = \theta(F_0)$ kojeg obično zovemo (dvostrani, simetrični) percentilni interval budući da je $\widehat{t_0}$ percentil distribucije od $|\theta(F_2) - \theta(F_1))|$ uvjetno na $F_1$. Interval sadrži $\theta_0$ s vjerojatnosti 0.95, što zovemo nominalna pokrivenost. Greška pokrivenosti se definira kao stvarna pokrivenost minus nominalna pokrivenost i obično konvergira u nulu kako se veličina uzorka povećava.

Sada ćemo detaljnije pogledati konstrukciju dvostranog, simetričnog, percentilnog intervala u parametarskim problemima. Pretpostavimo da su poznate funkcije distribucije $F_\lambda$ neprekidne. Tada se jednadžba (2.4) može egzaktno riješiti. Fokusirat ću se na slučajeve gdje je $\theta_0 = \theta(F_0)$ populacijska sredina, a populacija je normalna ili eksponencijalna.

Ako je populacija normalna $N(\mu, \sigma^2)$ i koristimo procjenitelj maksimalne vjerodostojnosti $\lambda = (\overline{X}, \widehat{\sigma}^2)$ za procjenu $\lambda_0 = (\mu, \sigma^2)$, onda uzoračku jednadžu (2.4) možemo zapisati kao

$$P(|n^{-1/2}\widehat{\sigma}N| \leq t|F_1) = 0.95, \tag{2.6}$$

gdje je $N$ normalna $N(0, 1)$ slučajna varijabla nezavisna od $X$. Slijedi da je rješenje od (2.4) $t = t_0 = x_{0.95}n^{-1/2}\widehat{\sigma}$, gdje je $x_\alpha$ definiran kao

$$P(|N| \leq x_\alpha) = \alpha.$$

Bootstrap pouzdani interval stoga izgleda ovako:

$$(\overline{X} - n^{-1/2} x_{0.95}\widehat{\sigma}, \overline{X} + n^{-1/2} x_{0.95}\widehat{\sigma}),$$

s greškom pokrivenosti

$$P(\overline{X} - n^{-1/2} x_{0.95}\widehat{\sigma} \leq \mu \leq \overline{X} + n^{-1/2} x_{0.95}\widehat{\sigma}) - 0.95 = P\{|n^{1/2}(\overline{X} - \mu)/\widehat{\sigma}| \leq x_{0.95}\} - 0.95. \quad (2.7)$$

Naravno, $n^{1/2}(\overline{X} - \mu)/\widehat{\sigma}$ nema normalnu distribuciju, nego reskaliranu (Studentovu) t-distribuciju s $n - 1$ stupnjem slobode.

Za drugi parametarski primjer uzet ću eksponencijalnu populaciju sa sredinom $\theta_0 = \mu$ i iskoristiti procjenitelj maksimalne vjerodostojnosti $\widehat{\lambda}$ za procjenu $\lambda_0 = \mu$. Onda je jednadžba (2.4)

$$P(\overline{X}|n^{-1}Y - 1| \leq t|F_1) = 0.95,$$

gdje $Y$ ima (gama) $\Gamma$ - distribuciju sa sredinom $n$ i nezavisna je od $X$. Rješenje uzoračke jednadžbe je tada $\widehat{t_0} = y_{0.95}\overline{X}$, gdje je $y_\alpha = y_\alpha(n)$ definirano s $P(|n^{-1}Y - 1| \leq y_\alpha) = \alpha$. Simetrični percentilni pouzdani intervala je tada

$$(\overline{X} - y_{0.95}\overline{X}, \overline{X} + y_{0.95}\overline{X}),$$

s greškom pokrivenosti

$$P(\overline{X} - y_{0.95}\overline{X} \leq \mu \leq \overline{X} + y_{0.95}\overline{X}) - 0.95 = P(|n^{-1}Y - 1| \leq y_{0.95}Y) - 0.95 = O(n^{-1}).$$

Općenito, simetrični percentilni pouzdani interval ima grešku pokrivenosti $O(n^{-1})$ (dokaz u 3.5 od [1]). Čak i simetrični, aproksimativni normalni interval (i u parametarskom i u neparametarskom slučaju), koji iznosi $(\overline{X} - n^{-1/2} x_{0.95}\widehat{\sigma}, \overline{X} + n^{-1/2} x_{0.95}\widehat{\sigma})$ kada je nepoznati parametar $\theta_0$ populacijska sredina, ima grešku pokrivenosti $O(n^{-1})$. Dakle nema ničeg posebno virtuoznog u vezi percentilnog intervala.

Kako bismo vidjeli zašto percentilni interval ima neadekvatnu izvedbu pogledajmo ponovno parametarski primjer s normalnom distribucijom. Uzrok problema je taj što se $\widehat{\sigma}$, a ne $\sigma$ pojavljuje na desnoj strani od (2.7). To je stoga što uzoračka jednadžba (2.4), ekvivalentna s (2.6), ovisi o $\widehat{\sigma}$. Drugim riječima, populacijska jednadžba (2.1) ekvivalentna s $P\{|\theta(F_1) - \theta(F_0)| \leq t\} = 0.95$, ovisi o $\sigma^2$, populacijskoj varijanci. Ovo se događa jer distribucija od $|\theta(F_1) - \theta(F_0)|$ ovisi o nepoznatom $\sigma$. Želimo eliminirati ili barem minimizirati tu ovisnost. Tu nam je od koristi pivotna funkcija. Egzaktnom pivotnom funkcijom osiguravamo da bootstrapom dobivamo pravu uzoračku distribuciju. Tada je greška pokrivenosti

0, za razliku od općenitog slučaja kada iznosi $O(n^{-1})$. Ako funkcija nije pivotna, bootstrap ne daje egzaktne zaključke.

Funkcija $T$ je (egzaktno) pivotna ako ima istu distribuciju za sve vrijednosti parametara, odnosno ne ovisi o njima. Asimptotski je pivotna ako, za nizove poznatih konstanti $\{a_n\}$ i $\{b_n\}$, $\{a_n\}T + \{b_n\}$ ima graničnu distribuciju koja ne ovisi o nepoznatim parametrima. Možemo pretvoriti $\theta(F_1) - \theta(F_0)$ u pivotnu statistiku ako ju promijenimo u $T = (\theta(F_1) - \theta(F_0))/\widehat{\tau}$, gdje je $\widehat{\tau} = \tau(F_1)$. $\widehat{\tau}$ može biti definirana na više načina, na primjer kao uzoračka standardna devijacija, $(n^{-1}\sum(X_i - \overline{X})^2)^{1/2}$, Gini-razlika sredina ili pak interkvartilni raspon. Gini-razlika sredina je mjera varijablinosti koja je informativnija u odnosu na varijancu za nenormalne distribucije. Za $X_i, i = 1, 2$ nezavisne i jednako distribuirane slučajne varijable $G_x$ deifinira kao: $G_x = E\{|X_1 - X_2|\}$ je Gini-razlika sredina od $X$. U kompleksnijim problemima, jackknife procjena standardne devijacije je obično opcija. Primijetimo da ćemo dobiti isti pouzdani interval ako zamijenimo $\widehat{\tau}$ s $c\widehat{\tau}$, za bilo koji $c \neq 0$, tako da nije nužno da je $\widehat{\tau}$ konzistentan s asimptotskom standardnom devijacijom od $\theta(F_1)$. Ono što je bitno je pivotnost - egzaktna pivotnost koja se ostvaruje ako imamo pouzdani interval s greškom pokrivenosti nula, odnosno asimptotska pivotnost ako takav interval nije moguće postići. Ako koristimo pivotnu statistiku $f_t$ se mijenja iz forme (2.3) u

$$f_t(F_0, F_1) = I\{\theta(F_1) - t\tau(F_1) \leq \theta(F_0) \leq \theta(F_1) + t\tau(F_1)\} - 0.95. \tag{2.8}$$

U ovom slučaju našeg parametarskog modela, bilo koji razumni procjenitelj $\widehat{\tau}$ će dati egzaktnu pivotnost. Uzmimo $\widehat{\tau} = \widehat{\sigma}$, gdje $\widehat{\sigma}^2 = \sigma^2(F_1) = n^{-1}\sum(X_i - \overline{X})^2$ označava uzoračku varijancu. Tada $f_t$ postaje

$$f_t(F_0, F_1) = I\{\theta(F_1) - t\sigma(F_1) \leq \theta(F_0) \leq \theta(F_1) + t\sigma(F_1)\} - 0.95. \tag{2.9}$$

Koristeći ovaj funkcional umjesto onoga u (2.3) sa normalnom populacijom, s potpuno istim argumentiranjem, jednadžba (2.6) mijenja se u

$$P\{(n-1)^{-1/2}|T_{n-1}| \leq t|F_1\} = 0.95, \tag{2.10}$$

gdje $T_{n-1}$ ima (Studentovu) t-distribuciju s $n-1$ stupnjeva slobode. (Stoga je uvjetovanje na $F_1$ u (2.10) irelevantno.) Prema tome, rješenje uzoračke jednadžbe je $\widehat{t_0} = (n-1)^{-1/2}w_{0.95}$, gdje je $w_\alpha = w_\alpha(n)$ dano s $P(|T_{n-1}| \leq w_\alpha) = \alpha$. Bootstrap pouzdani interval je $(\overline{X} - \widehat{t_0}\widehat{\sigma}, \overline{X} + \widehat{t_0}\widehat{\sigma})$.

Ovakvi pouzdani intervali se uobičajeno zovu *percentilni – t* intervali budući da je $\widehat{t_0}$ percentil statistike $|\theta(F_2) - \theta(F_1)|/\tau(F_2)$.

Na isti način moguće je konstruirati pouzdani interval za eksponencijalnu populaciju. Definirajmo $f_t$ kao u (2.9) i redefiniramo $w_\alpha$ s

$$P[|n^{-1/2}\sum_{i=1}^{n}(Z_i-1)|\{n^{-1}\sum_{i=1}^{n}(Z_i-\overline{Z})^2\}^{-1/2}\leq w_\alpha]=\alpha,$$

gdje su $Z_1,...,Z_n$ nezavisne eksponencijalne varijable s jediničnim očekivanjem. Rješenje $t_0$ uzoračke jednadžbe je $t_0 = w_{0.95}n^{-1/2}$ i $t-$ *percentilni* bootstrap pouzdani interval jest $(\overline{X}-\widehat{t_0}\widehat{\sigma},\overline{X}-\widehat{t_0}\widehat{\sigma})$.

Ovaj eksponencijalni primjer ilustrira slučaj u kojem $\widehat{\tau}=\widehat{\sigma}$ (uzoračka standardna devijacija) kao procjenitelj nije najprikladniji izbor. U ovim okolnostima, $\sigma(F_0)=\theta(F_0)$, čiji procjenitelj metodom maksimalne vjerodostojnosti jest $\theta(F_1)=\overline{X}$. Ovo sugerira uzimanje $\widehat{\tau}=\tau(F_1)=\overline{X}$ u definiciji (2.8) od $f_t$. Tada, ako je $w_\alpha = w_\alpha(n)$ redefiniran s

$$P\{|n^{-1}Y-1|(n^{-1}Y)^{-1}\leq w_\alpha\}=\alpha,$$

gdje je $Y\gamma-$distriburiana slučajna varijabla sa sredinom $n$, rješenje uzoračke jednadžbe je $\widehat{t_0}$=$w_{0.95}$. Rezultantni bootstrap interval je

$$(\overline{X}-w_{0.95}\overline{X},\overline{X}+w_{0.95}\overline{X}).$$

Ovaj primjer završit ću s opaskom na izračunu kritičnih točaka, kao što je $\widehat{v}_\alpha$, uniformnom Monte Carlo simulacijom (detaljnije o tome u Dodatku II. od [1]). Pretpostavimo da želimo izračunati rješenje $\widehat{v}_\alpha$ jednadžbe

$$P[\{\theta(F_2)-\theta(F_1)\}/\tau(F_2)\leq\widehat{v}_\alpha|F_1]=\alpha, \tag{2.11}$$

ili, preciznije, vrijednost

$$\widehat{v}_\alpha = inf\{x: P[\{\theta(F_2)-\theta(F_1)\}/\tau(F_2)\leq x|F_1]\geq\alpha\}.$$

Izaberimo cijele brojeve $B\geq 1$ i $1\leq\nu\leq B$ takve da $\nu/(B+1)=\alpha$. Na primjer, ako je $\alpha=0.95$, tada bismo mogli uzeti $(\nu,B)=(95,99)$ ili $(950,999)$. Uvjetno na $F_1$, generirimo $B$ ponovljenih uzoraka $\{X_b^*, 1\leq b\leq B\}$ s funkcijom distribucije $F_1$. U neparametarskom slučaju pišemo $F_{2,b}$ za empirijsku funkciju distribucije $X_b^*$. U parametarskom slučaju, gdje je populacijska funkcija distribucije $F_{\lambda_0}$ i $\lambda_0$ je vektor nepoznatih parametara, neka $\widehat{\lambda}$ i $\widehat{\lambda_b^*}$ označavaju procjene od $\lambda_0$ izračunate iz uzorka $X$ i ponovljenog uzorka $X_b^*$ redom. Stavimo $F_{2,b}=F_{(\widehat{\lambda_b^*})}$. Za oba slučaja (parametarski i neparametarski) definiramo

$$T_b^*=\{\theta(F_{2,b})-\theta(F_1)\}/\tau(F_{2,b})$$

i pišemo $T^*$ za generičku $T_b^*$. U ovoj notaciji 2.11 je ekvivalentno s $P(T^*\leq\widehat{v}_\alpha|X)=\alpha$. Neka je $\widehat{v}_{\alpha,B}$ $\nu$-ta najveća vrijednost u $T_b^*$. Tada $\widehat{v}_{\alpha,B}\to\widehat{v}_\alpha$ s vjerojatnosti jedan, uvjetno na $X$, s $B\to\infty$. Vrijednost $\widehat{v}_{\alpha,B}$ je Monte Carlo aproksimacija od $\widehat{v}_\alpha$.

## 2.2 Primjer s portfoliom (neparametarski bootstrap)

Primjer je preuzet iz [5]. Kao što sam već spomenula, implementacija bootstrapa u R-u sastoji se od dva koraka. Prvo kreiramo funkciju koja računa statistiku od interesa. Zatim koristimo *boot* funkciju za izvodjenje bootstrapa iz skupa podataka ponovljenim uzorkovanjem s mogućim ponavljanjima dovoljan broj puta. U ovom primjeru koristit ćemo $Portfolio$ skup podataka koji se nalazi u ISLR paketu. Želimo odrediti koja je najbolja investicijska strategija u jednostavnom modelu. Pretpostavimo da želimo uložiti fiksnu sumu novaca u dvije financijske imovine koje ostvaruju povrate $X$ i $Y$ redom, gdje su $X$ i $Y$ slučajni iznosi. Uložit ćemo udio $\alpha$ našeg novaca u imovinu prinosa $X$ i ostatak, odnosno $1-\alpha$ u imovinu prinosa $Y$. Kako postoji varijabilnost vezana uz povrate ovih imovina, želimo odabrati $\alpha$ koji će minimizirati ukupni rizik, ili varijancu, naših ulaganja. Drugim riječima, želimo minimizirati $Var(\alpha X + (1-\alpha)Y)$. Lako se pokaže da je vrijednost koja minimizira rizik dana s:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

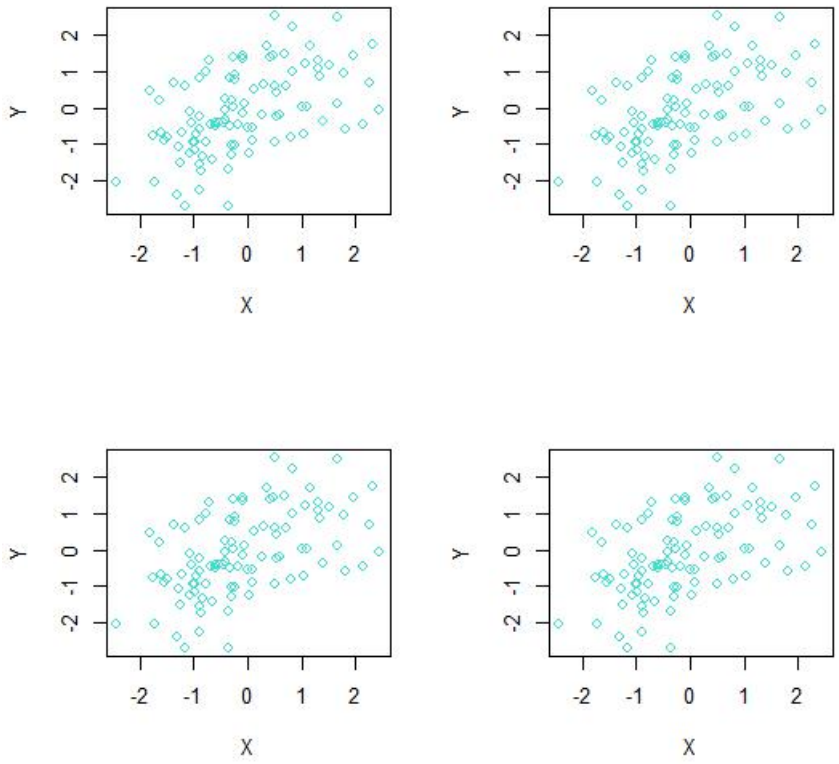
gdje $\sigma_X^2 = Var(X), \sigma_Y^2 = Var(Y)$ i $\sigma_{XY} = Cov(X, Y)$. U stvarnosti su iznosi $\sigma_X^2, \sigma_Y^2, \sigma_{XY}$ nepoznati. Možemo izračunati njihove procjene, redom $\widehat{\sigma}_X^2, \widehat{\sigma}_Y^2, \widehat{\sigma}_{XY}$, koristeći skup podataka koji sadrži prošle vrijednosti od $X$ i $Y$. Procijenimo vrijednost $\alpha$ koja minimizira varijancu našeg ulaganja koristeći:

$$\widehat{\alpha} = \frac{\widehat{\sigma}_Y^2 - \widehat{\sigma}_{XY}}{\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2 - 2\widehat{\sigma}_{XY}}. \tag{2.12}$$

Četiri puta simulirali smo 100 parova povrata na ulaganja (početni skup podataka je $Portfolio$) s povratima $X$ i $Y$. Zatim smo pomoću tih povrata procijenili $\widehat{\sigma}_X^2, \widehat{\sigma}_Y^2$ i $\widehat{\sigma}_{XY}$ koje smo potom uvrstili u (2.12) kako bismo dobili procjene za $\alpha$. Vrijednosti za $\widehat{\alpha}$ su u rasponu od 0.5752 do 0.6639. Slika 2.1 prikazuje simulirane parove.
Želimo kvantificirati točnost naše procjene. Za procjenu standardne devijacije od $\widehat{\alpha}$ ponavljamo 1000 puta proces simulacije 100 parova opservacija i procjene $\alpha$ (pomoću 2.12). Dobili smo 1000 procjena za $\alpha$ koje redom označimo kao $\widehat{\alpha}_1, \widehat{\alpha}_2, ..., \widehat{\alpha}_{1000}$. Na Slici 2.2 vidimo histogram dobivenih procjena. . Sredina za svih 1000 procjena za $\alpha$ iznosi:

$$\alpha = \frac{1}{1000}\sum_{r=1}^{1000}\widehat{\alpha}_r = 0.58,$$

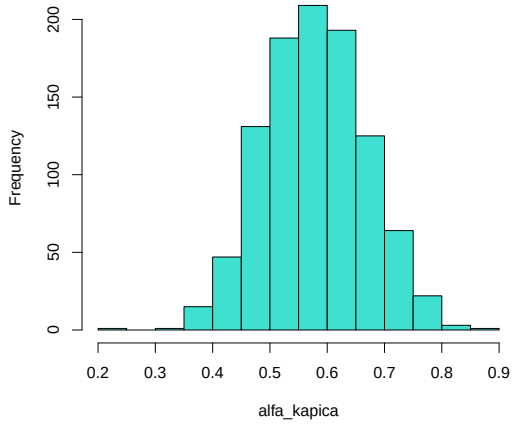
Standardna devijacija procjena je:




Slika 2.1: Četiri simulacije 100 parova povrata $(X, Y)$. Slijeva nadesno rezultat procjene za $\alpha$ je 0.6639, 0.5752, 0.6376, 0.6319

$$\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\widehat{\alpha}_r - \overline{\alpha})^2} = 0.088.$$

Ovo sugerira ideju o točnosti od $\widehat{\alpha}$ : $SE(\widehat{\alpha}) \approx 0.088$. Dakle, iz slučajnog uzorka iz populacije, očekujemo da se $\widehat{\alpha}$ razlikuje od $\alpha$ za približno 0.088, u prosjeku. U R-u se postupak jednostavno automatizira *boot* funkcijom. Slijedi grafički prikaz dobivenih procjena.
Imitirali smo proces generiranja uzoraka iz originalne populacije pomoću simulacija. Na taj način procijenili smo varijabilnost od $\widehat{\alpha}$ bez generiranja dodatnih uzoraka. Proces je ilustriran na Slici 2.3 na jednostavnom skupu podataka $Z$, koji sadrži samo $n = 3$ opservacije. Na slučajan način izaberemo $n$ opservacija iz skupa podataka. Dobijemo prvi bootstrap skup podataka $Z^{*1}$. Ponovljeno uzorkovanje je s mogućim ponavljanjima, od-




Slika 2.2: Histogram $\widehat{\alpha}$ izračunatih iz 1000 bootstrap uzoraka

nosno ista opservacija može se pojaviti više nego jednom u botstrap skupu podataka. Sada koristimo $Z^{*1}$ za novu procjenu od $\alpha$, koju zovemo $\widehat{\alpha}_1^*$. Ovaj postupak ponavljamo $B$ puta, za neku veliku vrijednost od $B$. Dobijemo $B$ različitih bootstrap skupova podataka, $Z^{*1}, Z^{*2}, ..., Z^{*B}$ i $B$ odgovarajućih procjena od $\alpha$, redom $\widehat{\alpha}_1^*, \widehat{\alpha}_2^*, ..., \widehat{\alpha}_B^*$. Standardnu grešku ovih bootstrap procjena računamo koristeći formulu

$$SE_B(\widehat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\widehat{\alpha}_r^* - \frac{1}{B}\sum_{r'=1}^{B}\widehat{\alpha}_r^{*\prime})^2}.$$

Ovo je zapravo približna vrijednost standardne greške od $\alpha$ procijenjena iz originalnog skupa podataka.

## 2.3 Primjer parametarski bootstrap

Primjer je preuzet iz [10]. Generiramo opet pouzdane intervale. Neka je $X$ slučajan uzorak iz eksponencijalne distribucije s parametrom $\lambda$, to jest da je funkcija gustoće $f(x|\lambda) = \lambda e^{-\lambda x}$. MLE (procjenitelj metodom maksimalne vjerodostojnosti) $\widehat{\lambda} = 1/\overline{X}$ i njegova je asimptotska varijanca jednaka $\lambda^2/n$. Kako znamo da je MLE asimptotski normalno distribuiran, $(1-\alpha)100\%$ pouzdani interval za $\lambda$ formiramo pomoću asimptotske varijance:

$$[\widehat{\lambda} - z(1-\alpha/2)\widehat{\lambda}/\sqrt{n}, \widehat{\lambda} + z(1-\alpha/2)\widehat{\lambda}/\sqrt{n}].$$

Generiramo slučajan uzorak veličine 100 iz eksponencijalne distribucije s parametrom $\lambda = 3$. Aritmetička sredina našeg uzorka je 0.34 pa je prema tome MLE za $\lambda$ jednak 2.91. Konstruiramo intervale pomoću procjene varijance i kvantila standardne normalne distribucije. Dobiveni 95% pouzdani interval za $\lambda$ je $[2.34, 3.48]$.

Pretpostavimo sada da ne želimo raditi s MLE ili da ne želimo koristiti asimptotsku distribuciju za formiranje pouzdanih intervala. Koristit ćemo bootstrap za procjenu distribucije od $\widehat{\lambda}$ i kreirati bootstrap pouzdane intervale za $\lambda$.

Prvo kreiramo skup bootstrap procjena parametra generiranjem $B$ slučajnih uzoraka veličine $n = 100$ iz eksponencijalne distribucije s parametrom $\widehat{\lambda}$ i koristimo svaki od uzoraka za dobivanje nove procjene parametra modela, $\widehat{\lambda}^{(b)}$. Općenito ne znamo ništa o distribuciji od $\widehat{\lambda}^{(b)}$. Svejedno možemo izračunati kvantile $q^*(\alpha)$, gdje je $q^*(\alpha)$ iz empirijske funkcije distribucije od $\widehat{\lambda}^b$. Možemo pisati:

$$P(q^*(\alpha/2) \leq \widehat{\lambda}^b \leq q^*(1-\alpha/2)) = 1-\alpha.$$

Sada promotrimo distribuciju od $\widehat{\lambda}^{(b)} - \widehat{\lambda}$. Iz gornjeg izraza vidimo:

$$P(q^*(\alpha/2) - \widehat{\lambda} \leq \widehat{\lambda}^b - \widehat{\lambda} \leq q^*(1-\alpha/2) - \widehat{\lambda}) = 1-\alpha.$$

Dodatno, možemo procijeniti distribuciju od $\widehat{\lambda} - \lambda$ pomoću distribucije od $\widehat{\lambda}^{(b)} - \widehat{\lambda}$. Možemo reći:

$$P(q^*(\alpha/2) - \widehat{\lambda} \leq \widehat{\lambda}^b - \widehat{\lambda} \leq q^*(1-\alpha/2) - \widehat{\lambda}) \doteq P(q^*(\alpha/2) - \widehat{\lambda} \leq \widehat{\lambda} - \lambda \leq q^*(1-\alpha/2) - \widehat{\lambda}) = 1-\alpha.$$

Za $\lambda$ formiramo pouzdani interval pomoću:

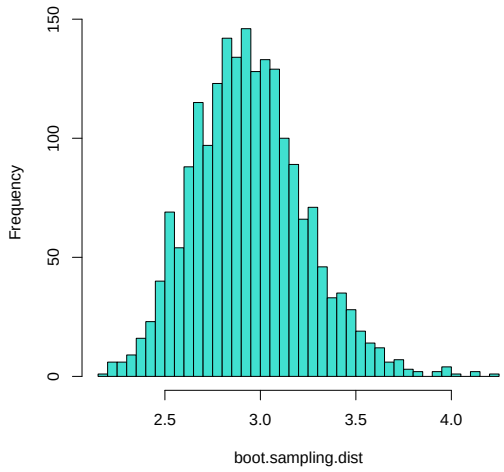
$$P(2\widehat{\lambda} - q^*(1-\alpha/2) \leq \lambda \leq 2\widehat{\lambda} - q^*(\alpha/2)) = 1-\alpha.$$

Slijedi da je bootstrap pouzdani interval za $\lambda$:

$$[2\widehat{\lambda} - q^*(1-\alpha/2), 2\widehat{\lambda} - q^*(\alpha/2)].$$

Postupak u R-u je sljedeći. Inicijaliziramo matricu koja će sadržavati vrijednost procjene za svaki uzorak (retčana matrica). Kreiramo 2000 bootstrap uzoraka i računamo vrijednost statistike za svaki uzorak. Na slici koja slijedi vidimo uzoračku distribuciju statistike.

Nakon izračuna kvantila distribucije računamo granice bootstrap pouzdanog intervala. Za naš slučajni uzorak $X$ dobiveni bootstrap 95% pouzdani interval za $\lambda$ je $[2.26, 3.39]$. Možemo izračunati i procjenu standardne greške od $\widehat{\lambda}$ pomoću uzoračke standardne greške bootstrap parametara. Vrijednost procjene u ovom slučaju je 0.29.




Slika 2.3: Uzoračka distribucija statistike

## 2.4 Primjer s nekoliko nezavisnih uzoraka

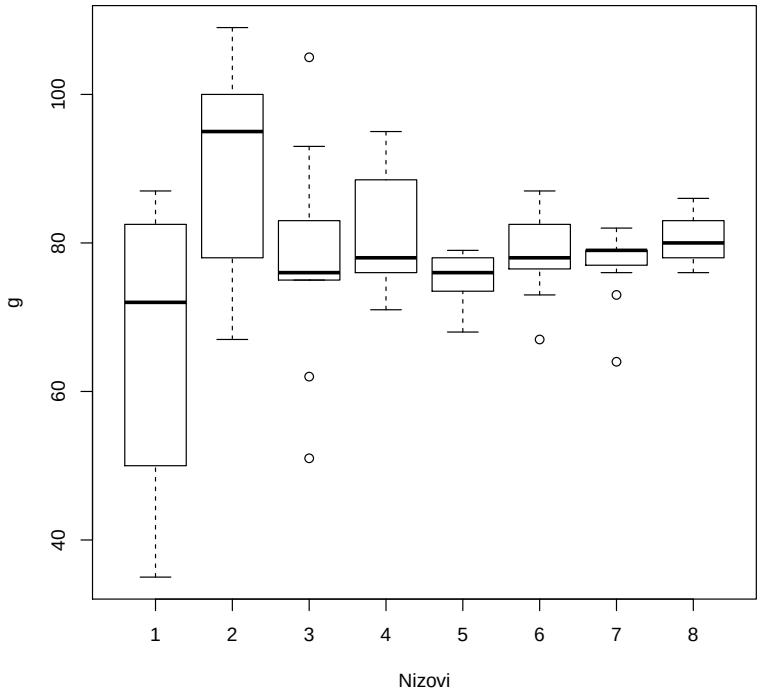
Primjer je preuzet iz [3]. Između svibnja 1934. i srpnja 1935. bio je izveden niz eksperimenata koji određuju akceleraciju uzrokovanu silom težom, $g$. Za potrebe eksperimenata bilo je provedeno osam mjerenja. Podaci za dobivene nizove su dani u sljedećoj tablici:

Tablica 2.1: Nizovi

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| 76 | 87 | 105 | 95 | 76 | 78 | 82 | 84 |
| 82 | 95 | 83 | 90 | 76 | 78 | 79 | 86 |
| 83 | 98 | 76 | 76 | 78 | 78 | 81 | 85 |
| 54 | 100 | 75 | 76 | 79 | 86 | 79 | 82 |
| 35 | 109 | 51 | 87 | 72 | 87 | 77 | 77 |
| 46 | 109 | 76 | 79 | 68 | 81 | 79 | 76 |
| 87 | 100 | 93 | 77 | 75 | 73 | 79 | 77 |
| 68 | 81 | 75 | 71 | 78 | 67 | 78 | 80 |
| | 75 | 62 | | | 75 | 79 | 83 |
| | 68 | | | | 82 | 82 | 81 |
| | 67 | | | | 83 | 76 | 78 |
| | | | | | | 73 | 78 |
| | | | | | | 64 | 78 |

Zanima nas parametar koji ovisi o populacijama $F_1, ..., F_8$ i skup podataka se sastoji od nezavisnih slučajnih uzoraka iz populacija. Tada je $i$-ti uzorak $y_{i1}, ..., y_{in_i}$ generiran iz populacije $F_i, i = 1, ..., 8$. Ako je $v = v(\widehat{F}_1, ..., \widehat{F}_8)$ procijenjena varijanca za $t$, pouzdani intervali za $\theta$ su bazirani na simuliranim vrijednostima od $z^* = (t^* - t)/v^{*1/2}$.

Graf koji slijedi nam sugerira da se varijanca smanjuje s rednim brojem uzorka (niza), te da postoji neznatna promjena u lokaciji, odnosno blagi outlieri su prisutni.




Slika 2.4: Box plotovi nizova ubrzanja sile teže

Pretpostavljamo da je svaki niz uzet iz zasebne populacije, $F_1, ..., F_8$, ali da svaka populacija ima sredinu $g$. Tada je primjereni procjenitelj težinska kombinacija

$$T = \frac{\sum_{i=1}^{8} (\mu(\widehat{F}_i)/\sigma^2(\widehat{F}_i)}{\sum_{i=1}^{8} (1)/\sigma^2(\widehat{F}_i)},$$

gdje je $\widehat{F}_i$ empirijska funkcija distribucije $i - tog$ niza, $\mu(\widehat{F}_i)$ procjena od $g$ iz $\widehat{F}_i$ i $\sigma^2(\widehat{F}_i)$ procijenjena varijanca za $\mu(\widehat{F}_i)$.

Procijenjena varijanca za $T$ je

$$V = \{\sum_{i=1}^{8}(1/\sigma^2(\widehat{F}_i))\}^{-1}.$$

Ako su podaci po pretpostavci normalno distribuirani sa sredinom $g$, ali različitim varijancama, trebali bismo uzeti

$$\mu(\widehat{F}_i) = \overline{y}_i, \sigma^2(\widehat{F}_i) = \{n_i(n_i - 1)\}^{-1} \sum_j (y_{ij} - \overline{y}_i)^2$$

za sredinu $i$-tog niza i njegovu procijenjenu varijancu. Rezultantni procjenitelj $T$ je tada empirijska verzija optimalne težinske sredine. Za naše podatke $t = 78.54$ sa standardnom greškom $v^{1/2} = 0.59$.

Sredina i varijanca od $T*$ su 78.51 i 0.371 tako da je procjena pristranosti za $T$ $78.51 - 78.54 \doteq -0.03$, a 95% pouzdani interval za $g$ baziran na normalnoj aproksimaciji (77.37, 79.76). $0.025 \cdot (R + 1)$ i $0.975 \cdot (R + 1)$ uređena statistika od $z_r^*$ su respektivno $-3.03$ i 2.50 tako da je 95% studentizirani bootstrap pouzdani interval za $g$ $(77.07, 80.32)$. Malo je širi od onoga koji je baziran na normalnoj aproksimaciji. S jačim pretostavkama o populaciji možemo se poslužiti na primjer poluparametarskim bootstrapom, o kojemu će biti riječi u sljedećem potpoglavlju. Prednost ovdje prikazanog postupka ponovljenog uzorkovanja je njegova robusnost.

## 2.5 Poluparametarski bootstrap

U poluparametarskom modelu neki aspekti distribucije su specificirani, dok su ostali proizvoljni. Kao jednostavan primjer može nam poslužiti karakterizacija $Y = \mu + \sigma\epsilon$, koji nema pretpostavke o distribuciji od $\epsilon$, osim da mu je centar nula i obujam jedan. Obično su poluparametarski modeli korisni jedino kada su podaci nehomogeni, s jedinom razlikom u parametrima. U kontekstu prethodno opisanog primjera, na primjer, možemo biti prilično sigurni da se distribucije $F_i$ razlikuju samo u obujmu i lokaciji. Odnosno, $Y_{ij}$ se može zapisati kao

$$Y_{ij} = \mu_i + \sigma_i\epsilon_{ij},$$

gdje je $\epsilon_{ij}$ uzorkovan iz zajedničke distribucije s kumulativnom funkcijom distribucije $F_0$. Normalna distribucija je parametarski model za ovu formu. Forma se donekle može provjeriti grafičkim prikazom standardiziranih reziduala, kao na primjer

$$e_{ij} = \frac{y_{ij} - \widehat{\mu_i}}{\widehat{\sigma_i}},$$

za odgovarajuće procjenitelje $\widehat{\mu_i}$ i $\widehat{\sigma_i}$ u svrhu provjere homogenosti pojedinog uzorka. Zajednička $F_0$ je procijenjena empirijskom funkcijom distribucije svih $\sum(n_i)$ od svih $e_{ij}s$, ili bolje empirijskom funkcijom distribucije standardiziranih reziduala $e_{ij}/(1 - n_i^{-1})^{1/2}$. Algoritam ponovnog uzorkovanja bi tada bio

$$Y^*_{ij} = \widehat{\mu_i} + \widehat{\sigma_i}\epsilon^*_{ij},\ j = 1, ..., n, i = 1, ..., k,$$

gdje su $\epsilon^*_{ij}$ slučajno reuzorkovane iz empirijske funkcije distribucije, odnosno slučajno ponovno uzorkovane sa zamjenom iz standardiziranih reziduala.

U drugom kontekstu, s pozitivnim podacima kao na primjer očekivana životna dob, možemo razmišljati o distribucijama na način da se one razlikuju samo u multiplikativnom efektu, odnosno da vrijedi $Y_{ij} = \mu_i\epsilon_{ij}$, gdje su $e_{ij}$ slučajno uzorkovane iz neke osnovne distribucije s jediničnom sredinom. Eksponencijalna distribucija je parametarski model za ovakve forme. Princip je uglavnom isti: procjenjujemo $e_{ij}$ s rezidualima $e_{ij} = y_{ij}/\widehat{\mu_i}$ i zatim definiramo $Y^*_{ij} = \widehat{\mu_i}\epsilon^*_{ij}$, s $\epsilon^*_{ij}$ slučajno ponovno uzorkovane sa zamjenom iz standardiziranih reziduala. Slična ideja primjenjuje se i u slučaju regresije, o kojoj ću u sljedećem poglavlju reći nešto više.

Ovakve metode ponovljenog uzorkovanja daju točnije rezultate u slučaju da su pretpostavke o $F_i$ točne, ali nisu robusne.

## 2.6 Smanjenje pogreške i opravdanost bootstrapa

Kao što sam već spomenula, u metodama ponovljenog uzorkovanja pogreška je općenito kombinacija statističke pogreške i pogreške simulacije. Prva je uzrokovana razlikom između $F$ i $\widehat{F}$ pa će veličina rezultantne greške ovisiti o odabiru statistike $T$. Greška simulacije je posljedica korištenja empirijskih procjenitelja svojstava kod reuzorkovanja iz $\widehat{F}$, umjesto egzaktnih svojstava.

### Statistička pogreška

Znamo da je ideja osnovnog bootstrapa aproksimirati neki iznos $c(F)$, kao na primjer $var(T|F)$ procjeniteljem $c(\widehat{F})$, gdje je $\widehat{F}$ parametarska ili neparametarska procjena od $F$ bazirana na uzorku $y_1, ..., y_n$. Statistička greška je tada razlika između $c(\widehat{F})$ i $c(F)$. Mi ju naravno želimo maksimalno minimizirati ili, ako je to moguće, eliminirati. Budući da je $T$ statistika o čijem izboru sami odlučujemo, $c(F)$ je kvantil ili neki moment iznosa $Q = q(\widehat{F}, F)$ dobivenog iz $T$, kao na primjer $h(T) - h(\theta)$ ili $(T - \theta)/V^{1/2}$, gdje je $V$ procijenjena varijanca ili nešto kompliciranije kao omjer vjerodostojnosti. Statistički problem u ovom slučaju je odabir između mogućih iznosa tako da je rezultantni $Q$ pivotni koliko je moguće, odnosno da ima (barem aproksimativno) istu distribuciju kod reuzorkovanja iz $F$ i $\widehat{F}$.

Pod pretpostavkom da je $Q$ monotona funkcija od $\theta$ možemo neposredno dobiti granice pouzdanosti. Na primjer, ako je $Q = h(T) - h(\theta)$ sa $h(T)$ rastućim po $T$ i ako je $a_\alpha$ pripadni donji $\alpha$ kvantil od $h(T) - h(\theta)$, tada

$$1 - \alpha \doteq Pr\{h(T) - h(\theta) \geq a_\alpha\} = Pr[\theta \leq h^{-1}\{h(T) - a_\alpha\}], \tag{2.13}$$

gdje je $h^{-1}(.)$ inverzna transformacija. Dakle, $h^{-1}\{h(T) - a_\alpha\}$ je gornja $(1 - \alpha)$ pouzdana granica za $\theta$.

U parametarskim problemima vrijedi da $\widehat{F} \equiv F_{\widehat{\psi}}$ i $F \equiv F_\psi$ imaju istu formu, razlikujući se samo u vrijednostima parametara. Ideja pivota je u ovom slučaju prilično jednostavna, označavajući konstantno ponašanje po svim vrijednostima parametara modela. Formalnije, definiramo pivot funkciju kao $Q = q(T, Q)$ i proučavamo njeno ponašanje u modelu za razne vrijednosti parametara. Može se provjeriti da, ponekad u teoriji i uvijek empirijski, možemo simultano proučavati karakteristike od $T - \theta, logT - log\theta$ i njihovu studentiziranu

verziju simulacijom nekoliko eksponencijalnih modela koji su blizu prilagođenom modelu. Ovo bi moglo rezultirati usporedbom izabranih kvantila u odnosu na parametarske vrijednosti, iz čega možemo dijagnosticirati nepivotalno ponašanje od $T-\theta$ i pivotalno ponašanje od $logT - log\theta$. Posebna uloga za transformaciju $h(T)$ proizlazi iz činjenice da je ponekad relativno lagano izabrati $h(.)$ tako da je varijanca od $T$ aproksimativno ili egzaktno nezavisna od $\theta$ i ta stabilnost je primarno svojstvo stabilnosti distribucije. Pretpostavimo da $T$ ima varijancu $\nu(\theta)$. Tada pod uvjetom da se funkcija $h(\cdot)$ se dobro ponaša u okolini od $\theta$, razvoj Taylorovog reda oko $\theta$ vodi do

$$var\{h(T\} \doteq \{\dot{h}(\theta)\}^2 \nu(\theta), \tag{2.14}$$

gdje je $\dot{h}$ prva derivacija $dh(\theta)/d\theta$. Jednadžba (2.13) implicira da je varijanca učinjena aproksimativno konstantom (odnosno jednakom jedan) ako je

$$h(t) = \int^t \frac{du}{\{v(u)\}^{1/2}}. \tag{2.15}$$

Potonje je poznato pod nazivom transformacija stabilizacije varijance. Bilo koji višekratnik od $h(t)$ bit će jednako učinkovit: često u jednouzoračkim problemima gdje je varijanca definirana kao $v(\theta) = n^{-1}\sigma^2(\theta)$ jednadžba (2.14) bi sadržavala $\sigma(u)$ umjesto $\{v(u)\}^{1/2}$. U tom slučaju $h(\cdot)$ je nezavisan od $n$ i $var(T) \doteq n^{-1}$. Za probleme u kojima $v(\theta)$ jako varira s $\theta$, korištenje ove transformacije zajedno sa (2.12) tipično daje točnije granice pouzdanosti nego što bi bilo dobiveno korištenjem direktnih aproksimacija kvantila za $T-\theta$. Je li takvo korištenje aproksimacije primjereno ponekad će biti jasno iz teoretskih razmatranja, kao u eksponencijalnom slučaju.

U neparametarskom slučaju situacija je kompliciranija. Sada nije vjerojatno da je bilo koji iznos točno pivotan. Također, ne možemo simulirati podatke iz distribucije koja je iste forme kao i $F$ zato što je njena forma nepoznata. Ipak, možemo simulirati podatke iz distribucija koje su slične empirijskoj funkciji distribucije $\widehat{F}$ i to bi moglo biti dovoljno budući da je $F$ blizu $\widehat{F}$. Smanjenje statističke pogreške je u fokusu istraživanja metoda ponovljenog uzorkovanja. Ovo se posebice odražava u razvoju točnih metoda za određivanje granica pouzdanosti, koje su opisane u poglavlju 5 od [3]. Bootstrap metoda kojom je moguće smanjiti statističku grešku opisana je u poglavlju 3.9.1 od [3].

## Pogreška simulacije

Pogreška simulacije se javlja kada se izvode Monte Carlo simulacije i svojstva statistika su u njima aproksimirana njihovim empirijskim svojstvima. Na primjer, aproksimiramo procjenu $B = E^*(T^*|\widehat{F}) - t$ pristranosti $\beta = E(T) - \theta$ sredinom $B_R = R^{-1}\sum(T_r^* - t) = \overline{T}^* - t$

koristeći nezavisne replikacije $T_1^*, ..., T_R^*$, svaku baziranu na slučajnom uzorku iz $\widehat{F}$. Monte Carlo varijabilnost u $R^{-1}\sum(T_r^*)$ može biti u potpunosti uklonjena jedino beskonačnom simulacijom, koja je naravno nemoguća i nepotrebna u praksi. Pitanje koje nam je od interesa glasi, koliko veliki treba biti $R$ kako bi se ostvarila razumna točnost, relativna statističkoj točnosti iznosa (pristranost, varijanca itd.), aproksimirana simulacijom? Nije moguće dati općenit odgovor, ali možemo prilično dobro odrediti što je potrebno uzimajući u obzir pristranost, varijancu i procjene kvantila u jednostavnim slučajevima. To ću sada i napraviti.

Pretpostavimo da imamo uzorak $y_1, ..., y_n$ iz $N(\mu, \sigma^2)$ distribucije i da je parametar od interesa $\theta = \mu$ procijenjen uzoračkom sredinom $t = \bar{y}$. Koristimo neparametarsku simulaciju za aproksimaciju pristranosti, varijance i $p$ kvantila $a_p$ od $T - \theta = \overline{Y} - \mu$. Korak koji slijedi je uzimanje $R$ nezavisnih repliciranih uzoraka iz $y_1, ..., y_n$ i računanje njihovih sredina $\overline{Y}_1^*, ..., \overline{Y}_R^*$. Iz ovoga računamo pristranost, varijancu i procjenitelje kvantila. Naravno, ovaj problem je prilično jednostavan pa su i očita rješenja $0, n^{-1}\sigma^2, n^{-1/2}\sigma z_p$, gdje je $z_p$ $p$-ti kvantil standardne normalne distribucije. Odgovarajuće procjene pristranosti i varijance su 0 i $n^{-1}\widehat{\sigma}^2$, gdje je $\widehat{\sigma}^2 = n^{-1}\sum(y_j - \bar{y})^2$. Odgovarajuća procjena $\widehat{a}_p$ od $p$ kvantila $a_p$ jest približno $n^{-1/2}\widehat{\sigma}z_p$, zanemarujući $O(n^{-1})$ uvjete. Sada ćemo usporediti aproksimacije dobivene konačnom simulacijom s ovim procjenama. Za početak promotrimo procjenitelj pristranosti.

$$B_R = R^{-1}\sum(\overline{Y}_r^* - \widehat{Y}).$$

Uvjetno na specifični uzorak $y_1, ..., y_n$ ili, ekvivalentno, na njegovu empirijsku funkciju distribucije $\widehat{F}$, sredina i varijanca procjenitelja pristranosti u svim mogućim simulacijama su

$$E^*(R^{-1}\sum(\overline{Y}_r^*) - \bar{y}) = 0, \quad var^*(R^{-1}\sum(\overline{Y}_r^*) - \bar{y}) = \frac{\widehat{\sigma}^2}{R_n}, \tag{2.16}$$

zato što je $E^*(\overline{Y}_r^*) = \bar{y}$ i $var^*(\overline{Y}_r^*) = n^{-1}\widehat{\sigma}^2$. Bezuvjetna varijanca od $B_R$, uzimajući u obzir varijabilnost između uzoraka od ishodišne distribucije je

$$var(R^{-1}\sum(\overline{Y}_r^*) - \bar{y}) = var_Y\{E^*(R^{-1}\sum(\overline{Y}_r^*) - \bar{y})\} + E_Y\{var^*(R^{-1}\sum(\overline{Y}_r^*) - \bar{y})\},$$

gdje $E_Y(.)$ i $var_Y(.)$ označavaju sredinu i varijancu uzetu u skladu sa zajedničkom distribucijom od $Y_1, ..., Y_n$. Iz (2.15) ovo daje

$$var(B_R) = var_Y(0) + E_Y(\frac{\widehat{\sigma}^2}{nR}) = \frac{\sigma^2}{n} \cdot \frac{n-1}{nR}. \tag{2.17}$$

Ovaj rezultat ne ovisi o normalnosti podataka. Sličan izraz vrijedi za bilo koju glatku statistiku $T$ s linearnom aproksimacijom.

Sada ćemo promotriti procjenitelj varijance $V_R = (R-1)^{-1}\sum(\overline{Y}_r^* - \overline{Y}_\cdot^*)^2$, gdje $\overline{Y}_\cdot^* = R^{-1}\sum \overline{Y}_r^*$. Sredina i varijanca od $V_R$ u svim mogućim simulacijama, uvjetno na podatke, su

$$E^*(V_R) = \frac{\widehat{\sigma}^2}{n}, \quad var^*(V_R) \doteq (\frac{\widehat{\sigma}^2}{n})^2\frac{2}{R}(1+\frac{1}{2}\widehat{\gamma}_2),$$

gdje je $\gamma_2$ standardizirana os za podatke (Appendix A u [3]). Primijetimo da bi $\gamma_2$ bio nula u parametarskoj simulaciji, ali ovdje ne (iako općenito $\widehat{\gamma}_2 = O(n^{-1})$ jer su podaci normalno distribuirani). Bezuvjetna varijanca od $V_R$, uprosječivanjem svih mogućih skupova podataka, je

$$var(V_R) \doteq var_Y(\frac{\widehat{\sigma}^2}{n}) + E_Y(\frac{\widehat{\sigma}^2}{n})^2)\frac{2}{R}(1+\frac{1}{2}\widehat{\gamma}_2),$$

što se reducira u

$$var(V_R) \doteq \frac{2\sigma^4}{n^3} + \frac{2}{R}(\frac{2\sigma^4}{n^3} + \frac{\sigma^4}{n^2}). \tag{2.18}$$

Prvi izraz na desnoj strani od (2.18) nastaje zbog varijacije podataka, drugi zbog varijacije simulacije. Implikacija jest da u svrhu dobivanja varijance simulacije u vrijednosti od 10%, potrebno je uzeti $R = 10n$.

Na kraju promotrimo procjenitelj $p$-og kvantila od $a_p$ za $\widehat{Y}-\mu$, što je $\widehat{a}_{p,R} = \overline{Y}^*_{((R+1)p)} - \overline{y}$ sa $\overline{Y}^*_{((R+1)p)}$ kao $(R+1)$-om uređenom statistikom simuliranih vrijednosti $\overline{Y}_1^*, ..., \overline{Y}_R^*$. Općeniti izračun svojstava simulacije od $\widehat{a}_{p,R}$ je kompliciran pa ćemo mi pojednostaviti pretpostavku da je $N(\overline{y}, n^{-1}\widehat{\sigma}^2)$ kao aproksimacija od $\overline{Y}^*$ egzaktna. S ovom pretpostavkom standardne karakteristike od uređene statistike daju

$$E^*(\widehat{a}_{p,R}) \doteq \widehat{a}_p = n^{-1/2}\widehat{\sigma}z_p,$$

i

$$var^*(\widehat{a}_{p,R}) \doteq \frac{p(1-p)}{Rg^2(\widehat{a}_p)} = \frac{2\pi p(1-p)\widehat{\sigma}^2 exp(z_p^2)}{nR}, \tag{2.19}$$

gdje je $g(.)$ gustoća od $\overline{Y}^* - \overline{Y}$ uvjetno na $\widehat{F}$. Bezuvjetna varijanca za sve skupove podataka se reducira na

$$var^*(\widehat{a}_{p,R}) \doteq \frac{\sigma^2}{n}\{\frac{z_p^2}{2n} + \frac{2\pi p(1-p)exp(z_p^2)}{R}\}. \tag{2.20}$$

# Poglavlje 3

# Linearni regresijski model

## 3.1 Regresijski modeli

Regresijska analiza jedna je od najvažnijih i najčešćih statističkih analiza. U njoj se proučava utjecaj kovarijate (prediktora) na varijablu odziva, odnosno utjecaj nezavisne varijable $X$ na zavisnu varijablu $Y$. Ja ću se u ovom radu osvrnuti samo na linearnu regresiju, u kojoj je sredina slučajnog odziva $Y$ za vrijednost prediktora $x$ jednaka $(x_1, ..., x_n)^T\beta$. Ovakav model još zahtjeva specifikaciju prirode slučajne varijacije, koju za nezavisne odzive predstavlja vrijednost $var(Y|x)$. Za parametarsku analizu još bismo trebali specificirati distribuciju od $Y$. Bez nje, dani model je poluparametarski. Za linearnu regresiju s normalnim standardnim greškama konstantne varijance, teorija najmanjih kvadrata predstavlja egzaktnu metodu za analizu. Međutim, kod generalizacija na nenormalne greške i nekonstantne varijance (ne vrijedi homoskedastičnost), egzaktne metode su rijetke i tada se uglavnom koriste aproksimacijske metode bazirane na linearnoj aproksimaciji procjenitelja i centralnom graničnom teoremu. Metode ponovljenog uzorkovanja u ovom slučaju osiguravaju točnije analize. Promatramo realne opservacije $Y_i = y_i$, gdje je

$$Y_i = g_i(\beta) + \epsilon_i, \quad i = 1, 2, ..., n. \tag{3.1}$$

Funkcije $g_i$ su poznatog oblika i obično ovise o nekoj od kovarijati $c_i$, dok je $\beta$ $p \times 1$ vektor nepoznatih parametara. $\epsilon_i$ su nezavisni i jednako distribuirani za neku distribuciju $F$ na $R^1$,

$$e_i \overset{n.j.d.}{\sim} F, \quad i = 1, 2, ..., n, \tag{3.2}$$

gdje je $F$ u nekom smislu centrirana oko nule. Za dani $n \times 1$ vektor $y = (y_1, y_2, ..., y_n)'$, procjenjujemo $\beta$ minimizacijom neke mjere udaljenosti $D(y, \eta)$ između $y$ i vektora prediktora $\eta(\beta) = (g_1(\beta), g_2(\beta), ..., g_n(\beta))'$,

$$\widehat{\beta} = Arg \min_{\beta} D(y, \eta(\beta)). \tag{3.3}$$

Najčešći izbor za $D$ jest $D(y, \eta) = \sum_{i=1}^{n}(y_i - \eta_i)^2$. Pretpostavimo da su modeli (3.1)−(3.3) prekomplicirani za standardnu analizu, odnosno trebamo i procjenu uzoračkih svojstava od $\widehat{\beta}$. Na primjer, za $g_i(\beta) = e^{c_i\beta}$, $F$ nepoznate distribucijske forme i $D(y, \eta) = \sum |y_i - \eta_i|$ bootstrap algoritam (3.1) − (3.3) bi se mogao modificirati na sljedeći način:

1. Konstruirajmo $\widehat{F}$ stavljajući masu $1/n$ svakom promatranom rezidualu,

$$\widehat{F} : masa \frac{1}{n} \quad na \quad \widehat{\epsilon_i} = y_i - g_i(\widehat{\beta}). \tag{3.4}$$

2. Konstruirajmo bootstrap skup podataka

$$Y_i^* = g_i(\widehat{\beta}) + \epsilon_i^*, \quad i = 1, 2, ..., n, \tag{3.5}$$

gdje su $\epsilon_i^*$ nezavisni i jednako distribuirani iz $\widehat{F}$ i računamo

$$\widehat{\beta^*} : Arg \min_{\beta} D(Y^*, \eta(\beta)). \tag{3.6}$$

3. Nezavisno ponovimo korak $2B$ puta, dobivajući pritom bootstrap replikacije $\widehat{\beta^{*1}}, \widehat{\beta^{*2}}, ..., \widehat{\beta^{*B}}$. Kao procjenu kovarijacijske matrice od $\widehat{\beta}$ možemo uzeti

$$\widehat{cov}(\widehat{\beta}) = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\beta}^{*b} - \widehat{\beta}^*)(\widehat{\beta}^{*b} - \widehat{\beta}^*)'. \tag{3.7}$$

Primjer (Linearna regresija)

Primjer je uzet iz [2]. Uobičajena situacija kod linearne regresije jest $g_i(\beta) = c_i\beta$, gdje je $c_i$ $1 \times p$ vektor poznatih kovarijanti i $D(y, \eta) = \sum(y_i - \eta_i)^2$. Neka je $C$ $n \times p$ matrica s vektorom $c_i$ kao $i$-tim retkom i neka vrijedi $G = C'C$. Pretpostavimo još da je prvi element u svakom $c_i$ jednak 1 i da je $G$ punog ranga $p$. U ovom slučaju možemo izračunati bootstrap vrijednost od (3.7) bez obveze Monte Carlo ponovljenog uzorkovanja.

Primijetimo da $\widehat{F}$ ima očekivanu vrijednost nula i varijancu $\widehat{\sigma}^2 = \sum_{i=1}^{n}(\widehat{\epsilon}_i^2/n)$ i da je $Y_i^* = c_i\widehat{\beta}+\epsilon_i^*$ standardni linearni model napisan drugačijom notacijom. Standardni linearni model pokazuje da vrijedi $\widehat{\beta}^* = G^{-1}C'Y^*$ i da je

$$\widehat{cov}(\widehat{\beta}^*) = \widehat{\sigma}^2 G^{-1}. \tag{3.8}$$

Drugim riječima, bootstrap daje standardnu procjenu kovarijance u slučaju linearne regresije, uz korištenje $\sum(\widehat{\epsilon}_i^2)/n$ umjesto $\sum(\widehat{\epsilon}_i^2)/(n-p)$ za procjenu $\widehat{\sigma}^2$. Algoritam (3.4)−(3.7) ovisi o $\widehat{F}$ koji je razumna procjena od $F$ i može dati lažno optimistične rezultate ako prilagođavamo preparametrizirane modele u nadi da ćemo naći jedan dobar. Kao primjer pogledajmo običnu polinomijalnu regresiju na realnom pravcu. Promatrani podaci su oblika $(t_1, y_1), (t_2, y_2), ..., (t_n, y_n)$, gdje je $t_i$ vrijednost prediktivne varijable za $y_i$. Ako je $n = 20$ i mi prilagodimo polinom 18. stupnja, $g_i(\beta) = c_i\beta$, gdje je $c_i = (1, t_i, t_i^2, ..., t_i^{18})$, onda je vjerojatno da će $\widehat{\sigma}^2 = \sum(\widehat{\epsilon}_i/20)$ biti jako mali i da će (3.7) dati preoptimističnu procjenu od $\text{Cov}(\widehat{\beta})$. U ovom slučaju problem se može smanjiti korištenjem nepristrane procjene od $\sigma^2$, $\sum(\widehat{\epsilon}_i^2)$, umjesto $\widehat{\sigma}^2$. Kao oprezniju alternativu za (3.4) − (3.6) možemo koristiti jednouzorački bootstrap.

## 3.2 Predviđanje u linearnoj regresiji

Linearna regresija se često koristi za predviđanje novih odziva od $Y_+$ kada je varijabla poticaja jednaka $x_+$. Htjeli bismo intervalno procijeniti $Y_+$. Granice pouzdanosti za sredinu odgovora $x_+^T\beta$ moguće je dobiti ponovljenim uzorkovanjem na isti način kao i granice pouzdanosti za pojedine koeficijente. Međutim, granice pouzdanosti za sam odziv $Y_+$, obično nazivane granice predviđanja, zahtjevaju dodatno reuzorkovanje za simulaciju varijacije od $Y_+$ oko $x_+^T\beta$. Pretpostavimo da predviđamo $Y_+ = x_+^T\beta + \epsilon_+$ i da je prediktor točke $\widehat{Y}_+ = x_+^T\widehat{\beta}$. Slučajna greška $\epsilon_+$ je po pretpostavci nezavisna od slučajnih grešaka $\epsilon_1, ..., \epsilon_n$ u promatranim odgovorima i zbog jednostavnosti uzimamo da su sve jednako distribuirane, posebno i da sve greške imaju iste varijance. Kako bismo dobili točnost točkovnog prediktora, možemo procijeniti distribuciju prediktivne greške

$$e = \widehat{Y}_+ - Y_+ = x_+^T\widehat{\beta} - (x_+^T\beta + \epsilon_+)$$

s distribucijom od

$$e^* = x_+^T\widehat{\beta} - (x_+^T\widehat{\beta} + \epsilon_+^*), \tag{3.9}$$

gdje je $\epsilon_+^*$ reuzrokovana iz $\widehat{F}$ i $\widehat{\beta}$ je simulirani vektor procjenitelja. Ovo pretpostavlja homoskedastičnost slučajne pogreške. Bezuvjetna svojstva prediktivne greške odgovaraju uprosječivanju u distribucijama kako $\epsilon_+$, tako i $\widehat{\beta}$, što radimo u simulaciji ponavljanjem (3.9) za svaki skup vrijednosti $\widehat{\beta}^*$. Dobivši modificirane reziduale $e_j$ iz prilagodbe podacima, algoritam generira $R$ skupova, svaki sa $M$ prediktora:

Za $r = 1, ..., R$

1. simuliramo odzive $y_r^*$ prema $Y_j^* = \widehat{\mu}_j + \epsilon_j^*, \quad j = 1, ..., n$, sa $\widehat{\mu}_j = \widehat{\beta}_0 + \widehat{\beta}_1 x_j^*$ i $\epsilon_j^*$ slučajno uzorkovano iz $\widehat{G}$.
2. dobijemo procjenu najmanjim kvadratima $\beta_r^* = (X^T X)^{-1} X^T y_r^*$; zatim
3. za $m = 1, ..., M$
   a) uzorak $\epsilon_{+,m}^*$ iz $e_1 - \overline{e}, ..., e_n - \overline{e}$ i
   b) računamo prediktivnu grešku $\sigma_{rm}^* = x_+^T \widehat{\beta}_r^* - (x_+^T \widehat{\beta} + \epsilon_{+,m}^*)$

Ovdje je dopušteno koristiti $M = 1$. Ključna stvar je da je $RM$ dovoljno velik za procjenu potrebnih karakteristika od $e^*$. Primijetimo, ako je potrebno više predikcija za razne $x_+$, tada se samo 3. korak ponavlja za svaki $x_+$. Srednja kvadratna prediktivna greška je procijenjena simulacijskom sredinom kvadrata greške $(RM)^{-1} \sum_{r,m} (e_{rm}^* - \overline{e}^*)^2$. Korisniji bi bio $(1 - 2\alpha)\%$ pouzdani interval za $Y_+$, za koji trebamo $\alpha$ i $(1 - 2\alpha)$ kvantile $a_\alpha$ i $a_{1-\alpha}$ od prediktivne greške $\delta$. Onda bi prediktivni interval imao granice

$$\widehat{y}_+ - a_{1-\alpha}, \widehat{y}_+ - a_\alpha.$$

Tada bi se egzaktni, ali nepoznati, kvantili procijenili empirijskim kvantilima $\delta^*$, čije uređene vrijednosti označavamo s $\delta_{(1)}^* \leq ... \leq \delta_{(RM)}^*$. Bootstrap prediktivni intervali su:

$$\widehat{y}_+ - \delta_{(RM+1)(1-\alpha)}^*, \widehat{y}_+ - \delta_{(RM+1)(\alpha)}^*, \tag{3.10}$$

gdje je $\widehat{y} = x_+^T \widehat{\beta}$. Ovo je zapravo analogno osnovnoj bootstrap metodi za pouzdane intervale. Nešto bolji pristup, koji podržava standardnu normalnu teoriju, je rad sa studentiziranom prediktivnom greškom

$$Z = \frac{\widehat{Y}_+ - Y_-}{S},$$

gdje je $S$ drugi korijen sredine kvadrata reziduala za linearnu regresiju. Odgovarajuće simulirane vrijednosti su $z_{rm}^* = \delta_{rm}^* / s_r^*$, sa $s_r^*$ izračunatim u koraku 2. algoritma. $\alpha$ i $(1 - \alpha)$ kvantili od $Z$ su procijenjeni s $z_{((RM+1)\alpha)}^*$ i $z_{((RM+1)(1-\alpha))}^*$ redom, gdje su $z_{(1)}^* \leq ... \leq z_{RM}^*$

uređenih $RM$ vrijednosti svih $z_s^*$. Tada je studentizirani bootstrap prediktivni interval za $Y_+$ jednak

$$\widehat{y}_+ - sz^*_{(RM+1)(1-\alpha)}, \widehat{y}_+ - sz^*_{(RM+1)(\alpha)}. \tag{3.11}$$

## 3.3 Primjeri implementacije

Neka linearni regresijski model ima formu $y_j = x_j^T\beta + \epsilon_j$, gdje su $(y_j, x_j)$ redom odziv i $p \times 1$ kovarijacijski vektor za $j$-ti odaziv $y_j$. Obično nas zanimaju pouzdani intervali za parametre, izbor kovarijata ili predikcija budućeg odziva $y_+$ za novi $x_+$. Dvije osnovne sheme ponovljenog uzorkovanja su:

- ponovljeno uzorkovanje parova (s ponavljanjima) $(y_1, x_1), ..., (y_n, x_n)$, s bootstrap podacima

  $$(y_1, x_1)^*, ..., (y_n, x_n)^*,$$

  nezavisno uzetim iz empirijske distribucije s jednakim vjerojatnostima $n^{-1}$ iz $(y_j, x_j)$ i

- ponovljeno uzorkovanje pogrešaka. Dobivši prilagođene vrijednosti $x_j^T\widehat{\beta}$ na slučajan način uzmemo $\epsilon_j^*$ iz centriranih reziduala $e_1, ..., e_n$ i stavimo

  $$y_j^* = x_j^T\widehat{\beta} + \epsilon_j^*, \quad j = 1, ..., n.$$

U slučaju ponovljenog uzorkovanja reuzorkovana matrica dizajna neće biti jednaka izvornoj. Za relativno velike skupove podataka ovo neće predstavljati problem, ali može imati značaja ako je $n$ mali ili ako nekoliko opservacija ima jaki utjecaj na neke aspekte dizajna. Ako je krivi model prilagođen i koristimo ovu shemu, trebala bi se dobiti ispravna mjere nesigurnosti, odnosno ponovljeno uzorkovanje je u ovom slučaju robusno. Druga shema je efikasnija od ponovljenog uzorkovanja parova ako je model ispravan, ali nije robusna u slučaju krivog modela. Stoga treba biti oprezan kod odluke o modelu i odabiru sheme. Obje sheme mogu biti stratificirane u slučaju nehomogenih podataka. U najekstremnijoj formi stratifikacije, nivo se sastoji samo od jednog reziduala. Ovo zovemo divlji bootstrap i koristi se u neparametarskoj regresiji. Primjeri koji slijede preuzeti su redom iz [3] i [5].

## Ponovljeno uzorkovanje parova

Model doživljenja (Efron, 1988.)

Procjena regresijskih koeficijenata metodom najmanjih kvadrata je poželjna kada su greške približno normalne u distribuciji i homoskedastične. Ako greške imaju outliere ili distribucije dugog repa (moguće zbog heteroskedastičnosti), navedena metoda nije efikasna. Slijedi da bismo prilikom svake regresijske analize trebali odrediti imaju li reziduali outliere i je li pretpostavka o normalnosti opravdana. Ako postojanje outliera ne utječe na regresijski model, onda se oni izostavljaju tijekom prilagodbe modela. Velike outliere treba svakako ukloniti iz finalne regresijske analize. Dva su razloga za to. Prvi je da su metoda prilagodbe otporne na outliere često ne baš efikasne. Drugi je da outlieri mogu poremetiti analizu metode kao što su najmanji kvadrati, koja nije otporna na njih. Na ponovljenom uzorkovanju koje se temelji na modelu outlier se može pojaviti u bilo kojoj od $n$ vrijednosti. Za ponovljeno uzorkovanje parova, outlieri se tada pojavljuju s varijablinim frekvencijama i tako čine bootstrap procjenu previše varijablinom. Efekti se mogu dijagnosticirati pomoću jaccknife-after-bootstrap grafa.

Podaci (u paketu *boot*) koje koristimo su postoci preživljavanja štakora uslijed raznih doza radijacije, s dva ili tri ponavljanja u svakoj dozi, dostupni su u sljedećoj tablici.

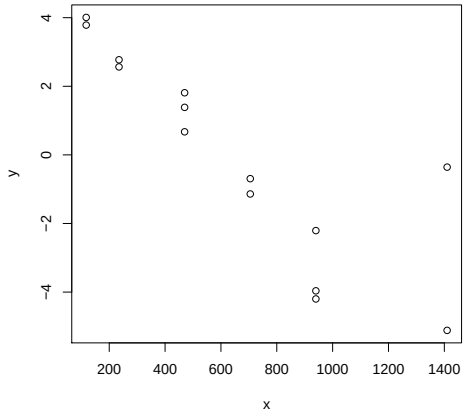
Tablica 3.1: Survival data

| Doza | 117.5 | 235.0 | 470.0 | 705.0 | 940.0 | 1410 |
|---|---|---|---|---|---|---|
| Postotak preživljenja | 44.0<br>55.0 | 16.0<br>13.0 | 4.0<br>1.96<br>6.120 | 0.5<br>0.32 | 0.11<br>0.015<br>0.019 | 0.7<br>0.006 |

Teoretska veza između stope preživljavanja i doze je eksponencijalna pa se linearna regresija primjenjuje na:

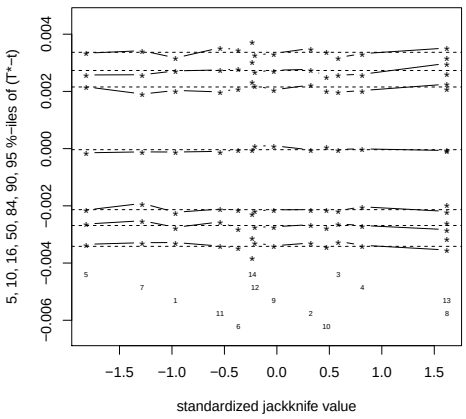
$$x = doza, y = log(postotak preživljenja).$$

Na slici koja slijedi vidimo navedene varijable. Prisutan je outlier za $n = 13$, $x = 1400$. Za ilustraciju djelovanja outliera u regresiji izvodimo ponovljeno uzorkovanje parova.




Slika 3.1: Graf podataka o preživljavanju

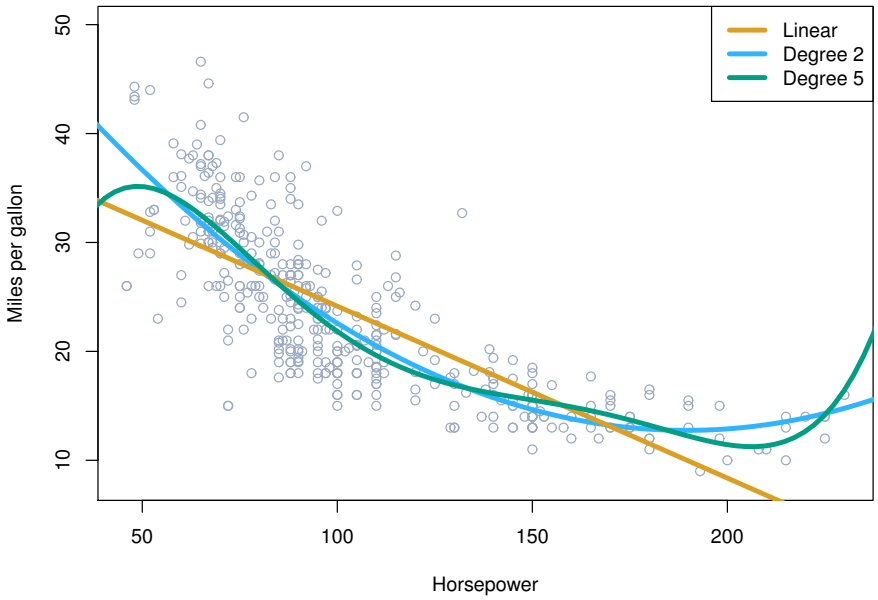
Procjena nagiba metodom najmanjih kvadrata je $-59\times10^{-4}$ ako koristimo sve podatke. Mijenja se u $-78\times10^{-4}$ sa standardnom greškom $5.4\times10^{-4}$ ako je podatak za $n = 13$ izostavljen. Iz grafa vidimo moguću heteroskedastičnost, stoga reuzorkujemo parove. Jackknife-after-bootstrap grafom pokazujemo efekt od $T^* - \widehat{t}$ reuzorkovanjem iz skupova podataka gdje je svaka od opservacija bila izuzeta. Ovdje očekujemo veliki utjecaj outliera. Efekt outliera na nagib možemo procijeniti i ubacivanjem *sim* = *parametric* u pozivu funkcije.




Slika 3.2: Jackknife-after-bootstrap graf za nagib

## Procjena točnosti linearnog regresijskog modela

Procjenjujemo varijabilnost parametara $\beta_0$ i $\beta_1$ linearnog regresijskog modela koristeći *Auto* skup podataka iz ISLR paketa. Skup se sastoji od 397 opservacija, koje među ostalim sadrže vrijednosti *mpg* i *horsepower*, koje redom označavaju potrošnju i snagu pojedinog automobila. Na slici vidimo kako postoji nelinearna veza među podacima.




Slika 3.3: Prikazani su *mpg* i *horsepower* za dani skup podataka *Auto*. Prilagodba linearnog modela označena je narančastom linijom. Prilagodba linearnog modela za model koji uključuje $horsepower^2$ je prikazana plavom linijom, a ona koja sadrži sve potencije od *horsepower* do 5. stupnja je prikazana zelenom linijom.

Općenito, jednostavni pristup za pripojenje nelinarosti linearnom modelu je uključivanje transformacijskih verzija prediktora u model. U našem slučaju mogao bi izgledati ovako:

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon.$$

Ovo je zapravo jednostavna linearna regresija s $X_1 = horsepower$, $X_2 = horsepower^2$. Koristmo stoga standardni linearni software za izračun procjene nepoznatih parametara (za ilustraciju primjenjujemo najjednostavniji model $mpg = \beta_0 + \beta_1 \times horsepower$).

Usporedit ćemo procjene dobivene bootstrapom s onima koje se izračunaju pomoću formula za $SE(\widehat{\beta_0})$ i $SE(\widehat{\beta_1})$:

$$SE(\widehat{\beta_0})^2 = \sigma^2\Big[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big], \quad SE(\widehat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{3.12}$$

Kreiramo prvo funkciju *boot.fn*(), koja uzima *Auto* podatke i skup indexa, a vraća procjene nepoznatih parametara. Tada primjenjujemo funkciju na svih 392 opservacije u svrhu izračuna procjene $SE(\widehat{\beta_0})$ i $SE(\widehat{\beta_1})$ za cijeli skup podataka. Funkciju možemo primijeniti i u svrhu kreiranja bootstrap procjena za nepoznate parametre slučajnim uzorkovanjem s mogućim ponavljanjima. Sada koristimo i *boot* funkciju za izračun 1000 standardnih grešaka bootstrap procjena za $SE(\widehat{\beta_0})$ i $SE(\widehat{\beta_1})$. Slijedi kako su izračunate vrijednosti $SE(\widehat{\beta_0}) = 0.86$ i $SE(\widehat{\beta_1}) = 0.0074$. Izračun standardnih grešaka možemo dobiti i koristeći naredbu *summary*. Standardne procjene grešaka $SE(\widehat{\beta_0})$ i $SE(\widehat{\beta_1})$ dobivene formulom (3.12) su redom 0.717 i 0.0064. Indiciraju li postojeće razlike problem s bootstrap procjenom? Zapravo suprotno. Formula (3.12) se oslanja na određene pretpostavke kao što su zavisnost o nepoznatom parametru $\sigma^2$. Procjenjujemo ga rezidualnom sumom kvadrata. Iako se formula za standardnu grešku ne oslanja na točnost linearnog modela, procjena za $\sigma^2$ to upravo čini. Među ovim podacima postoji nelinaerna veza kao što smo vidjeli na Slici 3.3. Reziduali iz linearnog modela su 'napuhani', pa tako je i $\sigma^2$. Također, standradne formule pretpostavljaju da su $x_i$ fiksni i da je sva varijablinost iz grešaka $\epsilon_i$. Bootstrap se ne oslanja niti na jednu od ovih pretpostavki i stoga vjerojatnije daje točniju procjenu standardnih grešaka od $\beta_0$ i $\beta_1$ od *summary* funkcije.

Kako kvadratni model osigurava dobru prilagodbu podacima (također vidimo na Slici 3.3), postoji bolja korespondencija između bootstrap procjena i standardnih procjena od $\beta_0$ i $\beta_1$.

# Poglavlje 4

# Zaključak

## 4.1 Kada je bootstrap primjenjiv

Krucijalno pitanje u primjeni bootstrapa je jesu li rezultati dobiveni s podacima na koje nailazimo u praksi pouzdani. Sugerira nam kako ponovljeni uzorci sami po sebi mogu reći kada i kako bi bootstrap izračun mogao podbaciti i kako bi trebao biti ispravljen da donese korisne odgovore (o tome više u (3.10) od [3]).

Drugo pitanje je: u kojim idealiziranim uvjetima procedura ponovljenog uzorkovanja ima rezultate koji su u nekom smislu matematički korektni? Odgovor na ovo pitanje uključuje asimptotski okvir u kojem veličina uzorka teži u beskonačnost. On služi kao zaštitna mreža, odnosno isključuje procedure koje nemaju prikladne karakteristike velikih uzoraka. Ipak, od esencijalne je važnosti znati prepoznati implikacije kada bi bootstrap mogao podbaciti.

Opišimo teoretsku bazu za proceduru jednostavnim pojmovima. Neka imamo slučajni uzorak $Y_1, ..., Y_n$ ili ekvivalentno empirijsku funkciju distribucije $\widehat{F}$, iz koje želimo procijeniti svojstva neke standardne veličine $Q = q(Y_1, ..., Y_n; F)$. Na primjer, mogli bismo uzeti

$$Q(Y_1, ..., Y_n) = n^{1/2}\Big(\overline{Y} - \int y dF(y)\Big) = n^{1/2}(\overline{Y} - \theta).$$

Želimo procijeniti funkciju distribucije

$$G_{F,n}(q) = Pr\{Q(Y_1, ..., Y_n; F) \leq q|F\}, \tag{4.1}$$

gdje uvjetovanje na $F$ indicira da je $Y_1, ..., Y_n$ slučajni uzorak iz $F$. Bootstrap procjena od (4.1) je

$$G_{\widehat{F},n}(q) = Pr\{Q(Y_1^*, ..., Y_n^*; \widehat{F}) \leq q|\widehat{F}\}. \tag{4.2}$$

U ovom slučaju $Q(Y_1^*, ..., Y_n^*; \widehat{F}) = n^{1/2}(\overline{Y}^* - \overline{y})$. Da bi $G_{\widehat{F},n}$ bio proizvoljno blizu $G_{F,n}$, $n \to \infty$, tri uvjeta moraju biti zadovoljena. Pretpostavimo da se prava distribucija od $F$ nalazi u otvorenoj okolini $N$ u odgovarajućem prostoru distribucija i da se $\widehat{F}$ nalazi u toj okolini za gotovo sve $n$ s vjerojatnosti jedan. Tada su uvjeti sljedeći:

1. za bilo koji $A \in N$, $G_{A,n}$ mora slabo konvergirati limesu $G_{A,\infty}$;
2. ova konvergencija mora biti uniformna na $N$; i
3. funkcija sa $A$ u $G_{A,\infty}$ mora biti neprekidna.

Ovdje slaba konvergencija od $G_{A,n}$ u $G_{A,\infty}$ podrazumijeva da, s $n \to \infty$

$$\int h(u)dG_{A,n}(u) \to \int h(u)dG_{A,\infty}(u), \tag{4.3}$$

za sve integrabilne funkcije $h(\cdot)$ (vidi str. 38 u [3]). Pod ovim uvjetima bootstrap je konzistentan, odnosno za svaki $q$ i svaki $\epsilon > 0$, $Pr\{|G_{\widehat{F},n}(q) - G_{F,\infty}(q)| > \epsilon\} \to 0$, kako $n \to \infty$. Prvi uvjet osigurava postojanje granične vrijednosti od $G_{F,n}$ i bio bi potreban i u situaciji kada bi $\widehat{F}$ bila izjednačena s $F$ za $n \geq n'$, za neki $n'$. Kako $n$ raste, $\widehat{F}$ se mijenja pa su drugi i treći uvjet potrebni kako bi osigurali da se $G_{F,n}$ približi $G_{F,\infty}$ koliko je moguće svakim mogućim nizom od $\widehat{F}$. Ako bilo koji od ovih uvjeta nije zadovoljen, bootstrap bi mogao iznevjeriti.

## 4.2 Kada bootstrap podbacuje

Kao što sam već spomenula u uvodu, osnovna tri slučaja kada treba biti oprezan s odlučivanjem o primjeni bootstrapa su kada su naši podaci nepotpuni, zavisni ili korumpirani. Pretpostavljali smo do sada da je $F$ distribucija od interesa i da je potpuni uzorak $y_1, ..., y_n$ iz $F$ predmet našeg razmatranja. Ovo je važno zbog nekoliko stvari, na primjer

zbog garantiranja statističke konzistentnosti našeg procjenitelja $T$. Međutim, u nekim primjenama, opservacija koju dobijemo ponekad nije sami $y$. Na primjer, ako su naši podaci mjerenja na nizovima pacijenata moguće je da za određene pacijente neka mjerenja nisu mogla biti izvedena, odnosno da imamo nepotpune podatke za naš slučajni uzorak.

Metoda općenitog neparametarskog ponovljenog uzorkovanja nije ispravna za zavisne podatke. Ovo je moguće vrlo jednostavno ilustrirati u slučaju gdje je uzorak $y_1, ..., y_n$ iz jedne realizacije koreliranih vremenskih nizova. Na primjer, promotrimo uzoračku sredinu $\overline{y}$ i pretpostavimo da su podaci iz stacionarnog niza $\{Y_j\}$ čija je marginalna varijanca $\sigma^2 = var(Y_j)$ i čije su autokorelacije $\rho_h = corr(Y_j, Y_{j+h}),\ \ za\ \ h = 1, 2, ....$ Neparametarska bootstrap procjena za varijancu od $\overline{Y}$ je približno $s^2/n$ i za velike $n$ ovo se približava $\sigma^2/n$. Međutim, prava varijanca od $\overline{Y}$ jest

$$var(\overline{Y}) = \frac{\sigma^2}{n} \sum_{h=-(n-1)}^{n-1} (1 - \frac{|h|}{n})\rho_h. \tag{4.4}$$

U ovom slučaju bi se suma često prilično razlikovala od jedan i tada bi bootstrap procjena varijance bila dosta loša. Sličan problem se pojavljuje kod ostalih formi zavisnih podataka. Suština problema je da jednostavni bootstrap pretpostavlja da je njihova zajednička empirijska funkcija distribucije $F(y_1) \times ... \times F(y_n)$ i stoga reuzorkuje iz njene procjene $F(y_1)^* \times ... \times F(y_n^*)$. Što je netočno za zavisne podatke. Problem je što ne postoji očit način za procjenu zajedničke gustoće za $Y_1, ..., Y_n$ s danom jednom realizacijom. Za slabo zavisne podatke neparametarske metode ponovljenog uzorkovanja su prilično efikasne.

Što ako koristimo podatke koji imaju outliere? Ako su u uzorku nađeni veći outlieri oni bi trebali biti uklonjeni ili ispravljeni. Kada imamo parametarski model, outlier pomaže detektirati lošu prilagodbu modela. Kada pak nemamo parametarski model, odnosno $F$ je procijenjena empirijskom funkcijom distribucije, tada je važno detaljno analizirati dobivene vrijednosti simulacije kako bismo odredili jesu li zaključci ovisili krucijalno o nekim specifičnim opservacijama. Više o ovome možete pročitati u (3.10) od [3].

Ukratko bih ovo poglavlje svela na jednu rečenicu, Richard Hamming:

*The purpose of computing is insight, not numbers; garbage in, garbage out.*

# Bibliografija


[1] P. Hall, The Bootstrap and Edgeworth Expansion, Springer-Verlag, 1992.

[2] B. Efron, The Jackknife, the Bootstrap and Other Resampling Plans, Department of Statistics, Stanford University 1982.

[3] A. C. Davison, D. V. Hinkey, Bootstrap methods and their application, Cambridge University Press, 1997.

[4] Rand R. Wilcox, Introduction to robust estimation and hypothesis testing, Academic Press, 1997.

[5] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical learning with Applications in R, Springer Science+Business Media New York 2013

[6] A. C. Davison, D. Kuonen, *An introduction to the Bootstrap with applications in R*, dostupno na `http://www.statoo.com/en/publications/bootstrap_scgn_v131.pdf` (prosinac 2014.)

[7] H. Chen, *Bootstrap method*, dostupno na `http://www.math.ntu.edu.tw/~hchen/teaching/LargeSample/notes/notebootstrap.pdf` (siječanj 2015.)

[8] A. Robinson, icebreakeR, Department of Mathematics and Statistics, University of Melbourne, 2010.

[9] N. Coelho, *Bootstrap Example and Sample Code*, `https://www.stat.berkeley.edu/~nate/Stat135/Bootstrap.pdf` (siječanj 2015.)

[10] A. C. Davison, D. V. Hinkley and G. A. Young, Recent Developments in Bootstrap Methodology, Statistical Science, 2003., Vol. 18, No. 2, 141–157

[11] B. Efron, Computer-intensive methods in statistical regression, SIAM Review, 1988.

[12] S. Yitzhaki, Gini's Mean difference: a superior measure of variability for non-normal distributions, METRON - International Journal of Statistics, 2003.

# Sažetak

Cilj ovog diplomskog rada je dati kratki pregled bootstrap metode, posebice za linearni regresijski model. Metoda se koristi za procjenu nesigurnosti. Korisna je kod zaključivanja koje se bazira na kompleksnim postupcima za koje su teoretski rezultati nedostupni, nisu primjenjivi na dostupnim uzorcima zbog njihove veličine ili pak kada je dostupni uzorak mali. Također je primjenjiva kada nismo sigurni koji model primijeniti ili kada jednostavno trebamo 'quick and dirty' analizu.

Uvod sarži kratki opis metoda ponovljenog uzorkovanja (jaccknife, bootstrap, kros - validacija i randomizacijski test), glavnu ideju bootstrapa i motivacijski primjer (procjena modificirane sredine populacije).

U prvom poglavlju detaljnije se prikazani neparametarski, parametarski i poluparametarski bootstrap s odgovarajućim primjerima. Argumentirane su opravdanost bootstrapa i mogućnost smanjenja pogreške tijekom primjene metode.

U drugom poglavlju pokazana je primjena metode na primjeru linearnog regresijskog modela. Opisane su metode ponovljenog uzorkovanja parova i ponovljenog uzorkovanja pogrešaka. Dani su također primjeri.

Na kraju je sažeto kada je bootstrap primjenjiv, a kada podbacuje.

# Summary

The goal of this master's thesis is to give an introduction to the bootstrap (bootstrapping), especially for the linear regression model. The method is used for assessing uncertainty. It is useful when inference is to be based on a complex procedure for which theoretical results are unavailable or not useful for the sample sizes met in practice, where a standard model is suspect but it is unclear with what to replace it, or when 'quick and dirty' answer is required. It can also be used to verify the usefulness of standard approximations for parametric models, and to improve them if they seem to give inadequate inferences.

In the introduction you can find the description of the resampling methods (the jacckinife, the bootstrap, cross-validation and randomization test), the main principle of the bootstrap and a motivation example (estimate of the trimmed mean of the population).

In the first chapter parametric, nonparametric and semiparametric bootstrap are explained with the given examples. The validity of the method and reducing error are argumented.

In the second chapter the application on the regression model is shown. The methods of resampling the residuals and resampling the errors are described. The methods the examples are also given.

In the end there is a conclusion about when to use bootstrap and when it is not recommended.

# Životopis

Rođena sam 05. veljače 1988. u Zagrebu. Godine 1994. upisala sam Osnovnu školu Vladimira Nazora u Zagrebu. Godine 2006., nakon završetka XV. gimnazije, upisala sam preddiplomski studij Matematika na Prirodoslovno-matematičkom fakultetu, Matematički odsjek, u Zagrebu. Nakon završetka preddiplomskog studija 2010. godine upisujem Diplomski studij Financijske i poslovne matematike.