Bootstrapové odhady v jazyku R

Patrik Mihalech[1]

Abstrakt
Bootstrapové odhady sú založené na princípe náhodných výberov s opakovaním. Ide o výpočtovo intenzívnu metódu, ktorá predstavuje alternatívu k tradičným spôsobom odhadu neznámych parametrov a predovšetkým intervalov spoľahlivosti. Technika bootstrapu umožňuje odhad empirických výberových rozdelení takmer akejkoľvek štatistiky za použitia náhodných výberov. Okrem odhadu neznámych parametrov sa dá postup uplatniť aj na štatistiky regresných modelov ako je napríklad koeficient determinácie a určiť jeho intervaly spoľahlivosti. Použitie bootstrapových odhadov je vhodné predovšetkým v prípadoch, keď je analytické vyjadrenie skúmaných štatistík veľmi náročné alebo ani nie je možné. Vzhľadom k vysokej výpočtovej náročnosti je na tvorbu bootstrapových odhadov potrebné použiť výpočtovú techniku a vhodný štatistický softvér. Cieľom tohto článku je oboznámiť čitateľa s teoretickými postupmi tvorby bootstrapových odhadov a následne s ich kalkuláciou v prostredí programovacieho jazyka R, za použitia balíčka boot, ktorého výhoda spočíva predovšetkým vo vysokej flexibilite a v priamom odhade intervalových odhadov bez nutnosti ďalšieho programovania.

Kľúčové slová
neparametrický bootstrap, parametrický bootstrap, bodový odhad, intervaly spoľahlivosti, programovací jazyk R

Abstract
Bootstrap estimates are based on principle of resampling with replacement. It is a compute-intensive method that presents an alternative to traditional means of unknown parameter estimation and especially their confidence intervals. Bootstrap technique allows us to estimate sampling distribution of almost any statistic based on random sampling. Besides random parameter estimation, advance can be used also in regression model statistics estimation such as coefficient of determination to compute its standard error and confidence intervals. Usage of bootstrap estimates is appropriate especially in cases when analytical solution of statistics of interest is very difficult or is not possible at all. Given high compute demands, it is essential to use appropriate statistical software for the calculation. The aim of this article is to acquaint reader with theoretical advances of bootstrap estimates creation and subsequently their calculation by usage of programming language R. More specifically, package boot, which advance is high flexibility and direct computation of confidence interval estimation without any necessity for further programming.

Key words
non-parametric bootstrap, parametric bootstrap, point estimate, confidence intervals, programming language R

JEL classification
C13, C14, C63

[1] Ekonomická univerzita v Bratislave, Fakulta hospodárskej informatiky, Katedra štatistiky, Dolnozemská cesta 1, 852 35 Bratislava, patrik.mihalech@euba.sk.

1 Úvod

V súvislosti s vývojom výpočtovej techniky sa čoraz viac dostáva do popredia metóda bootstrapu ako alternatíva k tradičným postupom štatistickej indukcie. Bootstrap je výpočtovo intenzívna metóda na odhad rôznych štatistických metrík. Ide o metódu, ktorá je založená na náhodnom výbere s opakovaním. Táto technika umožňuje odhadovať výberové distribučné funkcie takmer akýchkoľvek štatistík za použitia metód náhodného výberu a to aj takých štatistík, pri ktorých to iné postupy štatistickej indukcie neumožňujú. Pojem bootstrap ako prvý použil Bradley Efron vo svojej práci o jackknife výberoch (Efron, 1979). Technika bootstrapu môže byť použitá na odhad štandardnej odchýlky akejkoľvek štatistiky a na získanie jej intervalov spoľahlivosti. Bootstrap je veľmi výhodný predovšetkým v prípade, keď sa intervaly spoľahlivosti analyticky vyjadriť nedajú alebo ich analytické vyjadrenie je veľmi zložité. Princíp bootstrapu sa dá použiť tiež na odhad regresných koeficientov, ale aj intervalov spoľahlivosti, napríklad pre koeficient korelácie alebo koeficient determinácie regresného modelu.

Práve kvôli vysokej flexibilite a relatívnej jednoduchosti sú bootstrapové odhady čoraz viac využívané v praxi. Za nevýhodu tejto metódy môžeme považovať vysokú výpočtovú náročnosť, a preto je táto metóda bez výpočtovej techniky prakticky nepoužiteľná. Z toho dôvodu vznikali v štatistických softvéroch špecializované balíčky („*package*“) priamo pre tento účel, aby užívateľom uľahčili prácu s bootstrap algoritmami. Niekoľko takýchto balíčkov vzniklo aj pre programovací jazyk R (Davison a Hinkley, 1996), ktorému sa budeme venovať v tomto článku. Jedným z nich je balíček *bootstrap*, ktorý vytvorili Efron a Tibshirani v roku 1993 (Tibshirani a Leisch, 2017) a druhý je balíček *boot*, ktorý naprogramoval A. J. Canty (Canty a Ripley, 2017). Z týchto dvoch balíčkov je viac používaný balíček *boot*, a preto cieľom tohto článku je oboznámiť čitateľa s teoretickými postupmi tvorby bootsprapových odhadov a následne s ich kalkuláciou v prostredí R práve pomocou tohto balíčka.

Princíp bootstrapu sa dá tiež rozšíriť a použiť aj v iných oblastiach, ako je napríklad analýza časových radov. Jednotlivým pozorovaniam môžeme priradiť váhy, a tak zabezpečiť aby pozorovania, ktoré sú novšie, boli vyberané do bootstrapových výberov častejšie ako tie, ktoré sú staršie. Prípadne je možné uplatniť postup pre blokový bootstrap (Künsch, 1989). Metódy založené na bootstrapových výberoch sú tiež populárne v oblasti strojového učenia. Bootstrapové agregovanie („*Bagging*“) sa používa napríklad pri rozhodovacích stromoch za účelom zníženia rozptylu modelu a vyvarovania sa overfittingu (Johnson a Kuhn, 2018).

2 Neparametrický (plný) bootstrap

Predpokladajme, že chceme robiť induktívne úsudky o parametri θ náhodnej premennej X, na základe údajov z výberového súboru $(x_1, x_2, \dots, x_n)$ s distribučnou funkciou $F(x; \theta)$. Induktívne úsudky sú založené na výberovom rozdelení odhadu $\hat{\theta}$. Výberové rozdelenie v tom prípade často získame na základe teoretických výsledkov.

Napríklad, ak predpokladáme, že výberový súbor $(x_1, x_2, \dots, x_n)$ sa riadi exponenciálnym rozdelením pravdepodobnosti s parametrom λ, na základe centrálnej limitnej vety vieme povedať, že náhodná premenná X má asymptoticky normálne rozdelenie $X \sim N(1/\lambda,\ 1/n\lambda^2)$, ktoré môžeme použiť pri odhade štatistík, intervaloch spoľahlivosti alebo v štatistických testoch o parametre λ. V praxi sa však môžeme stretnúť s prípadmi, keď predpoklady nie sú splnené alebo asymptotické výsledky nie sú vhodné a nechceme ich použiť, lebo výbery sú malé (Canty, 2002).

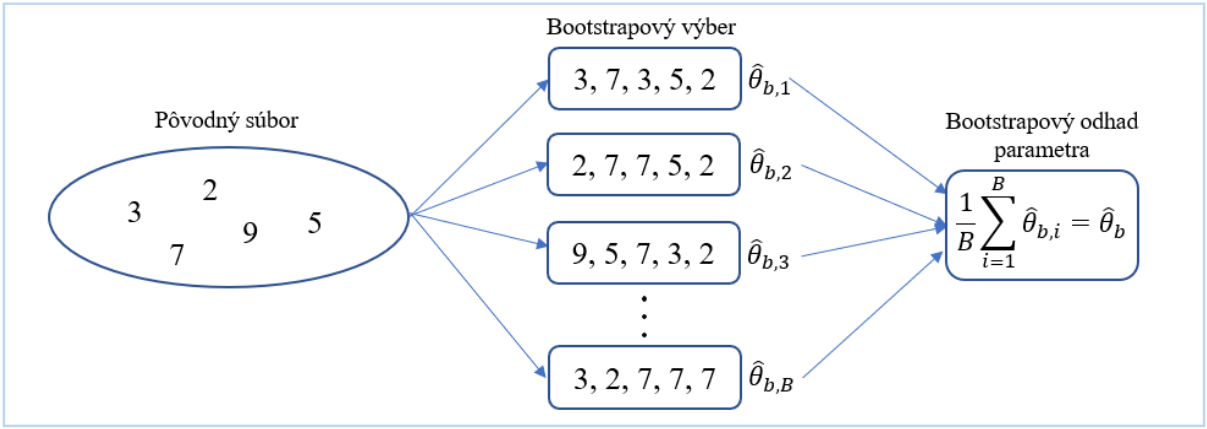
V tom prípade môžeme prijať induktívne úsudky založené na metóde bootstrapu. Bootstrap nám umožňuje vyhnúť sa úsudkom založeným na teoretickom výberovom rozdelení sledovanej štatistiky a miesto toho použiť empirické výberové rozdelenie. Toto dosiahneme opakovaným výberom z pôvodného súboru. V prípade, že nevychádzame z teoretického

rozdelenia pravdepodobnosti, ale priamo z údajov v štatistickom súbore, hovoríme o neparametrickom (plnom) bootstrape (Wasserman, 2010).

V prípade, že nepoznáme rozdelenie pravdepodobnosti náhodnej premennej X, nahradíme súbor pozorovaných hodnôt $(x_1, x_2, \dots, x_n)$ náhodného výberu $(X_1, X_2, \dots, X_n)$ novým súborom získaným z $(x_1, x_2, \dots, x_n)$ náhodným výberom s opakovaním. Takto získaný náhodný výber nazývame boostrapovým výberom. Pri tvorbe bootstrapových odhadov parametra θ náhodnej premennej X postupujeme nasledovne (Fox a Weissberg, 2018):

1. Z pozorovaných hodnôt $(x_1, x_2, \dots, x_n)$ náhodného výberu $(X_1, X_2, \dots, X_n)$ vypočítame odhad $\hat{\theta}$ parametra θ.
2. Následne realizujeme B náhodných bootstrapových výberov s opakovaním s rozsahom výberu n z pozorovaných hodnôt $(x_1, x_2, \dots, x_n)$. Odhad bude tým presnejší, čím vyšší bude počet bootstrapových odhadov. Vysoký počet odhadov však zvyšuje výpočtovú náročnosť. Vo všeobecnosti je vhodné zvoliť B prinajmenšom 1000.
3. Pre každý bootstrapový výber vypočítame odhad parametra θ a označíme ho $\hat{\theta}_{b,i}$, kde $i = 1, 2, \dots, B$.

Schéma 1: Postup pri odhade neznámeho parametra metódou bootstrap



Zdroj: vlastné spracovanie

Bootstrapovým odhadom vieme vypočítať predovšetkým nasledovné štatistiky (Derylo, 2018):

- Za bootstrapový odhad parametra θ obvykle považujeme aritmetický priemer.

$$\hat{\theta}_b = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i} \tag{1}$$

- Bootstrapový odhad rozptylu $\widehat{D}(\hat{\theta})$.

$$\widehat{D}(\hat{\theta})_b = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_{b,i} - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i}\right)^2 \tag{2}$$

- Bootstrapový odhad štandardnej odchýlky $\hat{s}(\hat{\theta})_b$.

$$\hat{s}(\hat{\theta})_b = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_{b,i} - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_{b,i}\right)^2} \tag{3}$$

- Bootstrapový odhad strednej kvadratickej chyby *MSE* odhadu $\hat{\theta}$.

$$\widehat{MSE}_b = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_{b,i} - \hat{\theta}\right)^2 \quad (4)$$

Bootstrapové odhady vytvoríme za použitia programovacieho jazyka R a porovnáme ich s inými metódami odhadu parametrov ako je napríklad metóda momentov. Na začiatku budeme pracovať s dátami vygenerovanými z exponenciálneho rozdelenia pravdepodobnosti s parametrom $\lambda = 0{,}35$. Kvôli reprodukovateľnosti výsledkov, si nastavíme hodnotu $seed = 38$. Na simuláciu dát použijeme funkciu `rexp` s rozsahom súboru 12, ktorú zapíšeme nasledovne:

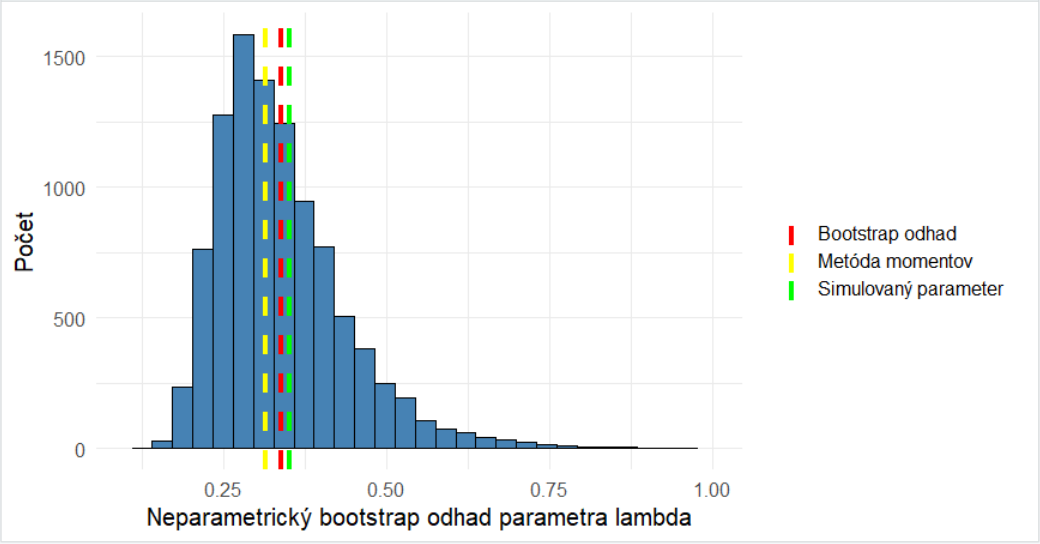
```
set.seed(38)
rexp(12, 0.35)
exp_data <- c(3.69, 10.75, 4.43, 3.43, 1.45, 1.43, 0.06, 2.07, 6.24, 0.05, 0.95, 3.66)
```

a výsledné dáta si uložíme na ďalšiu analýzu.

Metódou momentov vypočítame odhad parametra λ ako obrátenú strednú hodnotu $\hat{\lambda}_m = 1/E(X) \rightarrow \hat{\lambda}_m = 0{,}314$. Pri bootstrapovom odhade parametra $\hat{\lambda}_b$ budeme vychádzať zo vzťahu (1). Na náhodný výber zo súboru slúži v jazyku R funkcia `sample`. Pre tvorbu 10 000 náhodných výberov s opakovaním použijeme funkciu `replicate`, v rámci ktorej sa vypočíta parameter $\hat{\lambda}_{b,i}$ pre každý boostrapový výber samostatne. Po spriemerovaní týchto parametrov dostaneme výsledný boostrapový odhad $\hat{\lambda}_b = 0{,}338$. Zápis funkcie replicate a jej výsledok je nasledovný:

```
lambda_est <- replicate(10000, 1/mean(sample(exp_data, replace = TRUE)))
mean(lambda_est)
[1] 0.3383646
```

Graf 2:Histogram neparametrických bootstrapových výberov parametra exponenciálneho rozdelenia



Zdroj: vlastné spracovanie pomocou funkcie ggplot

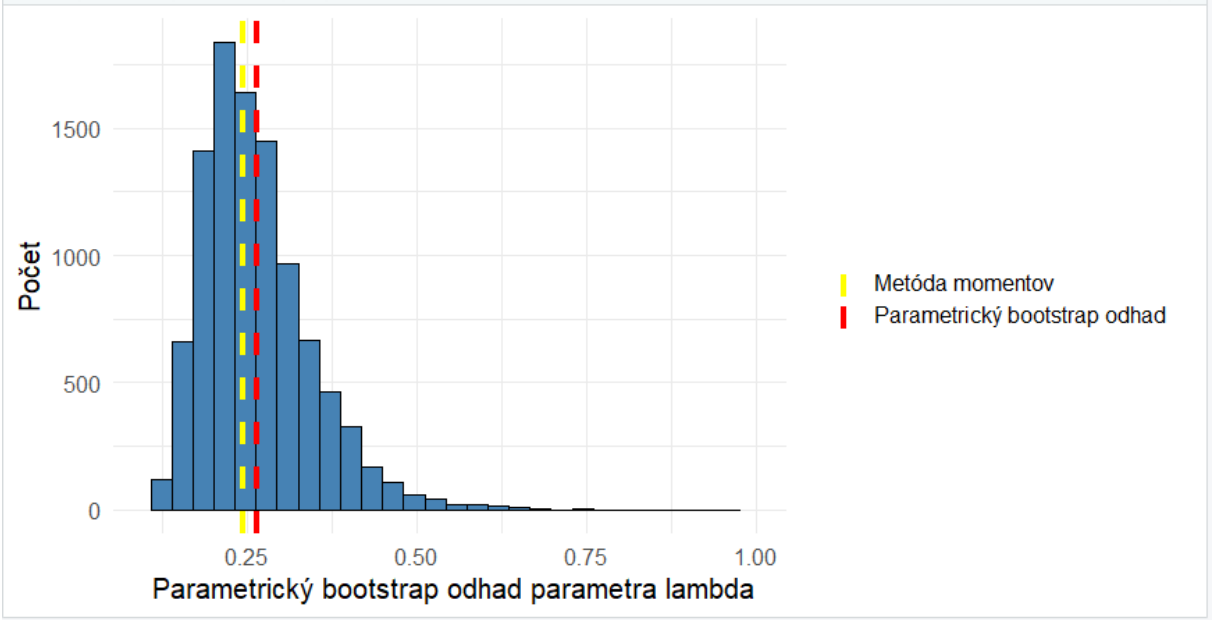
Vďaka jednotlivým odhadom z bootstrapových výberov môžeme vytvoriť histogram rozdelenia bootstrapových výberov. Bootstrapový odhad parametra je v porovnaní s momentovým odhadom bližšie skutočnému parametru (0,35), z ktorého boli dáta simulované. Relatívne veľký rozdiel medzi bootstrapovým a momentovým odhadom (0,024) je spôsobený malým súborom a tiež pravostranným zošikmením exponenciálneho rozdelenia. Na *Grafe 2* vidíme, že bootstrapové výbery sú zošikmené rovnako ako rozdelenie, z ktorého parameter odhadujeme.

3 Parametrický bootstrap

Okrem plného (neparametrického) bootstrapu, v niektorých prípadoch môžeme odhadovať neznáme štatistiky a intervaly spoľahlivosti aj parametrickým bootstrapom. Postup je veľmi podobný ako v prípade neparametrického bootstrapu. Jediný rozdiel je v tom, že bootstrapové výbery nevytvárame priamo z pôvodných údajov, ale najskôr parameter odhadneme a bootstrapové výbery potom vytvárame už z rozdelenia s daným parametrom. V tomto prípade si nasimulujeme dáta a vypočítame momentový odhad $\hat{\lambda}_m = 0{,}245$ a bootstrapové výbery realizujeme už priamo z exponenciálneho rozdelenia s parametrom 0,24. Príkaz v prostredí R bude nasledovný:

```
 set.seed(45)
 rexp(12, 0.35)
 param_exp_data <- c(0.76, 2.75, 4.66, 2.53, 4.33, 14.87, 3.34, 0.61, 2.79,
4.34, 3.79, 4.22)
 lambda_param_est <- replicate(10000, 1/mean(rexp(12,1/mean(param_exp_data)
)))
 mean(lambda_param_est)
[1] 0.2651123
```

Graf 2:Histogram parametrických bootstrapových výberov parametra exponenciálneho. rozdelenia



Zdroj: vlastné spracovanie pomocou funkcie ggplot

Vychýlenie medzi bootstrapovým a momentovým odhadom je v tomto prípade 0,0202. Rozdelenie parametrov vypočítaných z jednotlivých bootstrapových výberov je opäť

pravostranne zošikmené. Na základe bootstrapových rozdelení môžeme počítať k bootstrapovým odhadom aj intervaly spoľahlivosti. Práve intervaly spoľahlivosti sú najväčšou výhodou bootstrapových odhadov, lebo sa dajú počítať aj k štatistikám, ku ktorým sa štandardne intervaly spoľahlivosti počítať nedajú a je potrebné použiť rôzne aproximácie.

4 Výpočet bootstrapových odhadov a intervalov spoľahlivosti pomocou balíčka boot

Veľkou výhodou balíčka *boot* je, že priamo vypočíta k bootstrapovému odhadu aj intervaly spoľahlivosti. Základom je funkcia `boot`, ktorej povinné parametre sú dátový súbor, počítaná štatistika a počet bootstrapových výberov. Vzhľadom na to, že možnosti bootstrapových výberov sú veľmi široké, funkciu počítajúcu zvolenú štatistiku, si treba najskôr naprogramovať. Viac o programovaní užívateľských funkcií v prostredí jazyka R sa dá nájsť napríklad vo Wickham (2017) alebo Matloff (2011), pričom je potrebné mať aspoň základné vedomosti o syntaxe jazyka R. Výhodou funkcie `boot` je tiež to, že umožňuje počítať viac bootstrapových odhadov rôznych štatistík naraz.

Nasledujúci príkaz nám vytvorí zo simulovaných dát bodové odhady Pearsonovho koeficienta korelácie (*t1**), mediánu (*t2**) a aritmetického priemeru (*t3**). V základnom výstupe objektu vytvoreného prostredníctvom funkcie boot je originálny odhad z pôvodného súboru (*original*), odchýlka bootstrapového odhadu od pôvodného odhadu (*bias*) a prípustná chyba bootstrapového odhadu (*std. error*).

```
function_boot <- function(data, indices, corr.type){
  dt<-data[indices,]
  c(
    cor(dt[,1], dt[,2], method=corr.type),
    median(dt[,1]),
    mean(dt[,2])
  )
}
bootstrap_est <-boot(bootstrap_dataset, function_boot, R=1000, corr.type='p
')
bootstrap_est
```

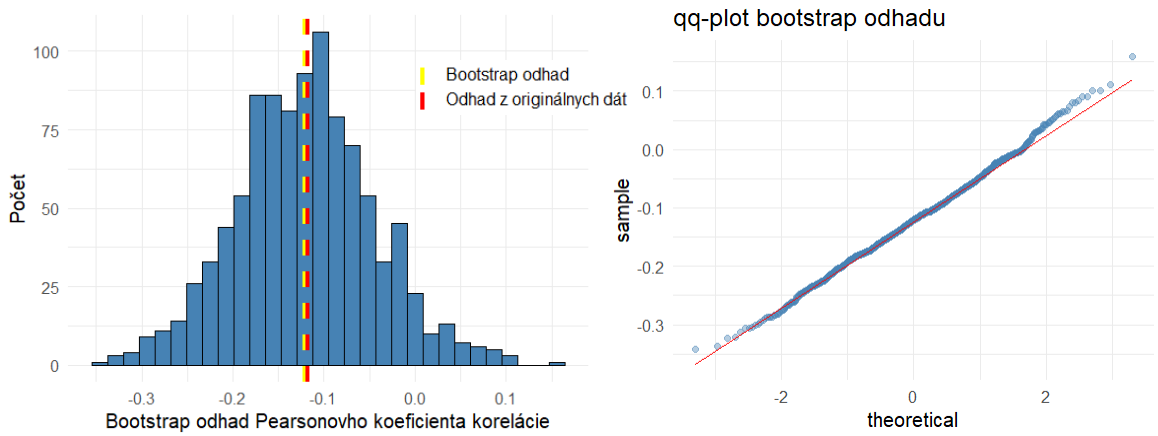
Obr.1: Výstup neparametrického bootstrapového odhadu získaného funkciou boot v programovacom jazyku R

```
ORDINARY NONPARAMETRIC BOOTSTRAP

          Bootstrap Statistics :
    original        bias    std. error
t1* -0.1175698  0.0013158343  0.07377321
t2*  5.8000000 -0.0134000000  0.09871770
t3*  3.0573333 -0.0006046667  0.03615418
```

Zdroj: Vlastné spracovanie v prostredí R

Graf 2: Histogram a qq-plot bootstrapového odhadu koeficienta korelácie



Zdroj: vlastné spracovanie pomocou funkcie ggplot

V jazyku R sa pomocou balíčka *boot* dá priamo počítať 5 typov intervalov spoľahlivosti pre bootstrapové odhady:

- Normálny interval spoľahlivosti (*norm*), kde sa pri výpočte používajú kvantily normálneho rozdelenia a najviac sa podobá na klasický interval spoľahlivosti, nepočítaný pomocou bootstrapových výberov. Rozdiel je v tom, že od bootstrapového odhadu ešte odpočítame vychýlenie (*bias*) od bodového odhadu z pôvodného súboru.

$$P(\hat{\theta}_b - bias - z_{1-\frac{\alpha}{2}} * s \leq \theta \leq \hat{\theta}_b - bias + z_{1-\frac{\alpha}{2}} * s) = 1 - \alpha \qquad (5)$$

  Normálny interval spoľahlivosti je prípustné použiť v prípade, ak je sledovaná štatistika neskreslená a má normálne rozdelenie. Avšak v prípade, že sú tieto podmienky splnené, väčšinou nie je potrebné vychádzať z bootstrapového odhadu, a preto sa vo všeobecnosti tento interval používa len málo.
- Percentilový interval spoľahlivosti (*perc*) sa počíta z kvantilov empirického rozdelenia boostrapového odhadu $\hat{\theta}_b$, a teda priamo na základe štatistík jednotlivých bootstrapových výberov.

$$P\left(\hat{\theta}_{b\frac{\alpha}{2}} \leq \theta \leq \hat{\theta}_{b_{1-\frac{\alpha}{2}}}\right) = 1 - \alpha \qquad (6)$$

  Percentilový interval spoľahlivosti sa môže použiť v prípade, že testovacia štatistika je neskreslená a homoskedastická. Tento interval je však vhodný iba ak je empirické bootstrapové rozdelenie symetrické. V opačnom prípade interval nemusí dávať správne výsledky. Oveľa robustnejšie riešenie pre rozdelenie s neštandardným tvarom je dosiahnuté za použitia základných bootstrapových intervalov.
- Základný interval spoľahlivosti (*basic*) vychádza z kvantilov empirického rozdelenia, tak ako je to pri percentilovom intervale spoľahlivosti definovanom vzťahom (6) , avšak v tomto prípade sa navyše využíva korekcia o vychýlenie medzi bootstrapovým odhadom a odhadom z pôvodného súboru $(\hat{\theta}_b - \hat{\theta})$. Po úprave dostaneme interval v tvare:

$$P\left(\hat{\theta}_{b\frac{\alpha}{2}} - \hat{\theta} \leq \hat{\theta}_b - \hat{\theta} \leq \hat{\theta}_{b_{1-\frac{\alpha}{2}}} - \hat{\theta}\right) = 1 - \alpha$$

$$P\left(2\hat{\theta} - \hat{\theta}_{b_{1-\frac{\alpha}{2}}} \leq \hat{\theta} \leq 2\hat{\theta} - \hat{\theta}_{b_{\frac{\alpha}{2}}}\right) = 1 - \alpha \quad (7)$$

Základné intervaly spoľahlivosti je vhodné použiť, keď je štatistika neskreslená a homoskedastická. V porovnaní s percentilovým intervalom spoľahlivosti dosahuje dobré výsledky aj pri štatistikách s neštandardným empirickým rozdelením odhadovaného parametra.

- Interval spoľahlivosti upravený o skreslenie (*BCa -bias corrected and accelerated*) je interval, ktorého výpočet vyžaduje používanie špeciálne upravených kvantilov bootstrapového rozdelenia. Kalkulácia *BCa* intervalu spoľahlivosti je matematicky pomerne zdĺhavá, postup výpočtu sa dá nájsť napríklad v DiCiccio a Efron, 1996. Tieto intervaly spoľahlivosti taktiež vyžadujú veľký počet bootstrapových výberov, lebo v opačnom prípade môžu viesť k nepresným výsledkom. Výhodou tejto metódy je, že pomáha redukovať skreslenie výsledných intervalov.
- Studentizovaný interval spoľahlivosti (*stud*) vychádza z náhodného výberu z bootstrapových výberov, ide teda o tzv. dvojstupňový bootstrapový odhad, ktorý sa použije na výpočet samotných intervalov spoľahlivosti. Studentizovaný interval spoľahlivosti nevie balíček *boot* priamo vypočítať, ale je potrebné najskôr spraviť úpravy, ktoré sú výpočtovo značne náročné. V studentizovaných intervaloch spoľahlivosti je pri malých výberových súboroch pomerne veľká prípustná chyba odhadu, čo spôsobuje, že tieto intervaly sú široké. Vzhľadom k vyššie uvedeným obmedzeniam sa studentizované bootstrapové intervaly v praxi veľmi nepoužívajú.

Na výpočet intervalov spoľahlivosti v jazyku R, slúži funkcia `boot.ci` (Páleš, 2017). Prvým argumentom je objekt `boot` – teda bootstrapový odhad, ktorý sme už v článku vypočítali skôr. Ďalším argumentom je typ bootstrapového intervalu (‘*norm*‘, ‘*basic*‘, ‘*perc*‘, ‘*bca*‘, ‘*stud*‘), spoľahlivosť odhadu a index prvku v objekte `boot`, pre ktorý chceme intervaly počítať. Pre 90, 95 a 99-percentné intervaly spoľahlivosti typu „basic“, “norm“, “perc“ a “bca“ pre Pearsonov koeficient korelácie sa príkaz boot.ci zapíše takto:

```
 boot.ci(boot.out = bootstrap_est, type = c('basic','norm','perc', 'bca'),
conf = c(.90, .95, .99), index = 1)
```

Jeho výsledkom je obr. 2. Poznamenajme, že studentizované intervaly spoľahlivosti sme nepočítali, pretože ich výpočet vyžaduje dvojstupňový bootstrapový odhad.

Obr. 2: Výstup intervalového bootstrapového odhadu pre koeficient korelácie v programovacom jazyku R

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

Intervals :
Level      Normal              Basic
90%   (-0.2394,  0.0051 )   (-0.2470, -0.0039 )
95%   (-0.2629,  0.0285 )   (-0.2719,  0.0178 )
99%   (-0.3086,  0.0743 )   (-0.3107,  0.0798 )

Level     Percentile            BCa
90%   (-0.2312,  0.0119 )   (-0.2227,  0.0226 )
95%   (-0.2530,  0.0368 )   (-0.2435,  0.0471 )
99%   (-0.3149,  0.0756 )   (-0.2875,  0.0990 )
Calculations and Intervals on Original Scale
```

Zdroj: Vlastné spracovanie v prostredí R

5 Bootstrap odhady v regresných modeloch

Na princípe bootstrapu vieme odhadnúť aj rôzne štatistiky regresných modelov. Takýto výpočet však môže byť, predovšetkým v prípade veľkého množstva bootstrapových výberov, časovo náročný. Ďalej uvádzame príklad na výpočet bootstrapových odhadov pre upravený koeficient determinácie (*t1**) a pre regresný koeficient (*t2**) modelu vytvoreného zo simulovaných dát. Funkcia `boot` musí pre každý bootstrapový výber vypočítať nový regresný model zo vstupných dát. V tomto prípade sme vytvorili 1000 regresných modelov, z ktorých sme si vždy uložili koeficient determinácie a regresný koeficient. Príkaz v prostredí R vyzerá nasledovne:

```
function_boot_lm <- function(data, indexy){
  dt<-data[indexy,]
  c(
    summary(lm(reg_y ~ reg_x, dt))$adj.r.squared,
    lm(reg_y ~ reg_x, dt)$coef[2]
  )
}
boot_reg <- boot(reg_data, function_boot_lm, R=1000)
boot_reg
```

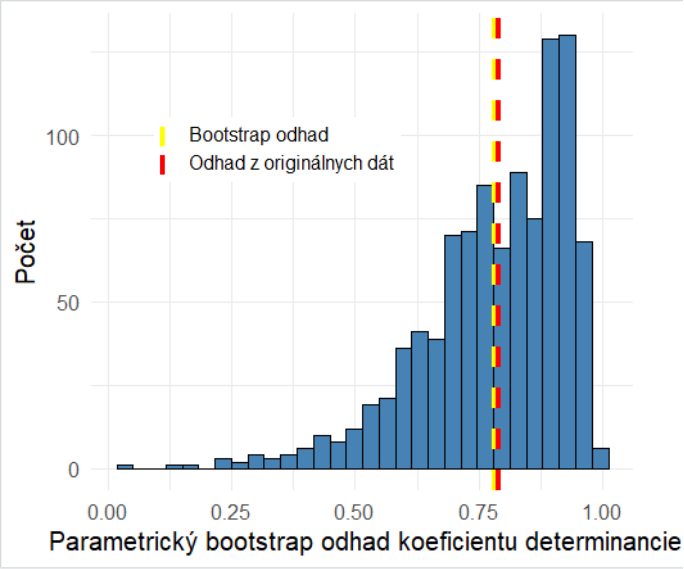
Obr. 3: Výstup bootstrapových odhadov pre koeficient determinácie a regresný koeficient v programovacom jazyku R

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Bootstrap Statistics :
      original        bias    std. error
t1*  0.79048133 -8.926988e-03 0.148668487
t2* -0.04496557  6.607597e-05 0.005951734
```

Zdroj: Vlastné spracovanie v prostredí R

Graf 2: Histogram odhadu koeficienta determinácie regresného modelu



Zdroj: vlastné spracovanie pomocou funkcie ggplot

Takýmto spôsobom vieme na základe bootstrapových výberov vypočítať intervaly spoľahlivosti aj pre koeficient determinácie. Zo zvolených bootstrapových intervalov sú v tomto prípade vhodnejšie percentilový a BCa interval, nakoľko horná hranica zvyšných intervalov je vyššia ako 1 (obr. 4). Príkaz pre výpočet intervalov spoľahlivosti pre koeficient determinácie vyzerá nasledovne:

```
boot.ci(boot_reg, type=c('basic', 'norm', 'perc', 'bca'),conf = .95, index
= 1)
```

Obr. 4: Bootstrapové intervalové odhady koeficientu determinácie v programovacom jazyku R

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap

Intervals :
Level      Normal              Basic
95%   ( 0.5080,  1.0908 )   ( 0.6157,  1.1663 )

Level     Percentile            BCa
95%   ( 0.4147,  0.9653 )   ( 0.2324,  0.9490 )
Calculations and Intervals on Original Scale
```

Zdroj: Vlastné spracovanie v prostredí R

Pre porovnanie môžeme uviesť, že za použitia percentilového bootstrapového intervalu dostaneme výsledok, ktorý je podobný intervalom vytvoreným za použitia Fisherovej transformácie (0,3499; 0,9370).

6 Záver

Bootstrapové odhady sa dostávajú v posledných desaťročiach vďaka výpočtovej technike čoraz viac do popredia. Hoci ich výpočet vyžaduje veľké množstvo kalkulácií, za pomoci výpočtovej techniky a štatistického softvéru je postup ich odhadu relatívne jednoduchý. Výhodou je ich priamočiarosť predovšetkým pri odhade štandardných odchýlok a intervalov spoľahlivosti pre štatistiky s komplexnejším rozdelením, ako sú napríklad korelačné koeficienty alebo pomery šancí v logistickej regresii. Bootstrap je tiež vhodným spôsobom ako skontrolovať stabilitu výsledkov a hoci vo väčšine prípadov nie je možné určiť skutočné intervaly spoľahlivosti, bootstrapové odhady sú asymptoticky presnejšie ako štandardné intervaly spoľahlivosti získané z výberového rozdelenia s predpokladom normality. Naopak, za nevýhodu bootstrapu by sme mohli považovať závislosť od pôvodného výberu a vyššiu časovú náročnosť.

V tomto článku sme ukázali ako je možné bez znalosti pokročilých programovacích techník vytvárať bootstrapové odhady a bootstrapové intervaly spoľahlivosti v programovacom jazyku R.

Literatúra

[1] Canty, A. (2002). Resampling Methods in R: The boot Package. *The newsletter of the R Project*.[online]. Dostupné na: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf [cit. 2021-02-01].

[2] Davison, A., Hinkley, D. (1996). *Bootstrap Methods and Their Application*. Cambridge University Press.

[3] DiCiccio T., Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science*, Vol. 11, No. 3 [online]. Dostupné na: https://projecteuclid.org/download/pdf_1 /euclid.ss/1032280214 [cit. 2021-02-01].

[4] Derylo, L. (2018). *Bootstrap in R* [online]. Dostupné na: https://www.datacamp.com /community/tutorials/bootstrap-r [cit. 2021-02-01].

[5] Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton.
[6] Fox, J., Weisberg, S. (2018). *An R Compation to Applied Regreesion: Bootstrapping Regression Models in R*. SAGE Publications, third edition.
[7] Johnson, K., Kuhn, M. (2018). *Applied Predictive Modeling*. Springer.
[8] Matloff, N. (2011). *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press, 1st edition.
[9] Páleš, M. (2017). *Jazyk R v aktuárskych analýzach*. Vydavateľstvo EKONÓM.
[10] R-bloggers. (2019). *Understanding Bootstrap Confidence Interval Output from the R boot Package* [online]. Dostupné na: https://www.r-bloggers.com/2019/09/understanding-bootstrap-confidence-interval-output-from-the-r-boot-package/ [cit. 2021-02-01].
[11] Ripley, B. (2020). *Package 'boot'*. The Comprehensive R Archive Network [online]. Dostupné na: https://cran.r-project.org/web/packages/boot/boot.pdf [cit. 2021-02-01].
[12] Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer.
[13] Wickham, H. (2017). *R for Data Science*. O'Reilly Media, 1st Edition.