Bootstrap metoda u linearnoj regresiji

Detić, Eleonora

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Eleonora Detić

BOOTSTRAP METODA U LINEARNOJ REGRESIJI

Diplomski rad

Voditelj rada:
doc. dr. sc. Snježana Lubura Strunjak

Zagreb, 2023.

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. _____

Mojim roditeljima i sestrama, za neizmjernu ljubav i podršku.

Hvala mentorici doc. dr. sc. Snježani Luburi Strunjak na pomoći oko pisanja ovog rada.
Hvala prijateljima, jer su uz vas dani bili kraći.
I na kraju, hvala L. Bilo je lakše udvoje.

Sadržaj

Uvod

Regresijska analiza, kao jedna od najčešćih metoda statističke analize, pronalazi široku primjenu u mnogim matematičkim problemima. Proučavajući utjecaj nezavisnih varijabli, tzv. kovarijata, donosi se zaključak o zavisnoj varijabli čija je distribucija od interesa. U ovom radu naglasak je na linearnoj regresiji, posebnom tipu regresijske analize, u kojoj je veza između zavisne i nezavisnih varijabli linearna.

Model linearne regresije potpuno je određen specificiranjem distribucije grešaka te njihovom varijancom. Ako se radi o modelu s normalno distribuiranim greškama i konstantnom varijancom, poznato je da metoda najmanjih kvadrata daje precizne, analitički točne rezultate. S druge strane, za generalizirane probleme u kojima greška nije normalna te varijanca nije konstantna, egzaktne matematičke metode gotovo da ne postoje. U takvim situacijama kao moguće rješenje nameću se aproksimacijske metode bazirane na centralnim graničnim teoremima - metode ponovljenog uzorkovanja.

U prvom poglavlju iznose se osnovni vjerojatnosni pojmovi te se opisuje *bootstrap* metoda. Drugo poglavlje definira jednostavan linearni model, objašnjava važnost pretpostavka te se opisuju mogućnosti metoda ponovnog uzorkovanja u slučaju kada te pretpostavke nisu zadovoljene. Treće poglavlje opisuje situaciju u višedimenzionalnom slučaju te se bavi prediktiranjem novih vrijednosti. Na kraju su navedeni praktični primjeri koji potkrepljuju teorijske rezultate. Svi primjeri potkrijepljeni su R kodom, koji je priložen kao dodatak na kraju rada.

Poglavlje 1

Uvod

1.1 Osnovni pojmovi teorije vjerojatnosti

Većina pojmova i definicija u ovom poglavlju preuzeta je iz [20].

Osnovni polazni objekt u teoriji vjerojatnosti jest neprazan skup Ω - prostor elementarnih događaja koji reprezentira skup svih mogućih ishoda nekog pokusa koji promatramo. Točke ω nazivamo elementarni događaji te odsad nadalje podrazumijevamo da je skup Ω zajedno sa svojim elementima, točkama ω, zadan.

Definicija 1.1.1. *Familija $\mathcal{F}$ podskupova od Ω ($\mathcal{F} \subseteq \mathcal{P}(\Omega)$) jest σ-algebra skupova (na Ω) ako je*

1. $\emptyset \in \mathcal{F}$
2. $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3. $A_i \in \mathcal{F}, i \in \mathbb{N} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Definicija 1.1.2. *Neka je $\mathcal{F}$ σ-algebra na skupu Ω. Uređen par $(\Omega, \mathcal{F})$ zove se izmjeriv prostor.*

Definicija 1.1.3. *Neka je $(\Omega, (F))$ izmjeriv prostor. Funkcija $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ jest vjerojatnost (na $\mathcal{F}$, na Ω) ako vrijedi:*

1. $\mathbb{P}(A) \geq 0, A \in \mathcal{F}$, $\mathbb{P}(\Omega) = 1$
2. $A_i \in \mathcal{F}, i \in \mathbb{N}, A_i \cap A_j = \emptyset$ *za* $i \neq j \implies \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

Definicija 1.1.4. *Uređena trojka $(\Omega, \mathcal{F}, \mathbb{P})$, pri čemu je $\mathcal{F}$ σ-algebra na Ω i $\mathbb{P}$ vjerojatnost na $\mathcal{F}$, zove se vjerojatnosni prostor.*

Budući da svakom rezultatu slučajnog pokusa pridružujemo neki realan broj, od interesa nam je promatrati realne funkcije na Ω. Borelovom σ-algebrom, $\mathscr{B}$, nazivamo σ-algebru generiranu svim otvorenim skupovima na $\mathbb{R}$, a njezine elemente zovemo Borelovi skupovi.

Definicija 1.1.5. *Neka je $(\Omega, \mathcal{F}, \mathbb{P})$ vjerojatnosni prostor. Funkcija $X : \Omega \to \mathbb{R}$ jest slučajna varijabla (na Ω) ako je $X^{-1}(B) \in \mathcal{F}$ za proizvoljno $B \in \mathscr{B}$, tj. $X^{-1}(\mathscr{B}) \subset \mathcal{F}$.*

U terminologiji jezika mjere slučajna varijabla je realna Borel-izmjeriva funkcija, tj. funkcija sa Ω u $\mathbb{R}$ izmjeriva u paru σ-algebri $(\mathcal{F}, \mathscr{B})$. Ipak, ključna razlika između teorije vjerojatnosti i teorije mjere je pojam funkcija distribucije.

Za $B \in \mathscr{B}$ stavimo

$$\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\{\omega \in \Omega, X(\omega) \in B\} = \mathbb{P}\{X \in B\}. \tag{1.1}$$

Relacijom 1.1 definirana je funkcija $\mathbb{P}_X : \mathscr{B} \to [0, 1]$, koju nazivamo vjerojatnosna mjera inducirana sa X, a vjerojatnosni prostor $(\mathbb{R}, \mathscr{B}, \mathbb{P}_X)$ nazivamo vjerojatnosni prostor induciran sa X.

Definicija 1.1.6. *Neka je X slučajna varijabla na Ω. Funkcija distribucije od X jest funkcija $F_x : \mathbb{R} \to [0, 1]$ definirana sa*

$$\begin{aligned} F_X(x) &= \mathbb{P}_X(\langle -\infty, x]) = \mathbb{P}(X^{-1}\langle -\infty, x]) = \mathbb{P}\{\omega \in \Omega; X(\omega) \leq x\} \\ &= \mathbb{P}\{X \leq x\}, \end{aligned} \tag{1.2}$$

za svaki $x \in \mathbb{R}$.

Ako je F funkcija distribucije, sa $C(F)$ označujemo skup svih točaka neprekidnosti od F.

Teorem 1.1.7. *Funkcija distribucije F slučajne varijable X je rastuća i neprekidna zdesna na $\mathbb{R}$ i zadovoljava*

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0, \tag{1.3}$$

$$F(+\infty) = \lim_{x \to +\infty} F(x) = 1. \tag{1.4}$$

Definicija 1.1.8. *Kažemo da slučajna varijabla X ima distribuciju teškog repa, tzv. heavy tailed distribuciju, ako vrijedi*

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty, \quad \text{za svaki } t > 0. \tag{1.5}$$

Definicija 1.5 preuzeta je iz [12].

Razlikujemo dva glavna tipa slučajnih varijabli: diskretne i neprekidne.

Definicija 1.1.9. *Slučajna varijabla X je diskretna ako postoji konačan ili prebrojiv skup* $D \subset \mathbb{R}$ *takav da je* $\mathbb{P}\{X \in D\} = 1$.

Definicija 1.1.10. *Neka je X slučajna varijabla na vjerojatnosnom prostoru* $(\Omega, \mathcal{F}, \mathbb{P})$ *i neka je* F_x *njezina funkcija distribucije. Kažemo da je X apsolutno neprekidna ili, kraće, neprekidna slučajna varijabla, ako postoji nenegativna realna Borelova funkcija f na* $\mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}_+$, *takva da je*

$$F_X(x) = \int_{-\infty}^{x} f(t)d\lambda(t) \tag{1.6}$$

za svaki $x \in \mathbb{R}$.

Integral u 1.6 je Lebesqueov integral funkcije f u odnosu na Lebesqueovu mjeru λ opisan u [5]. Funkciju iz 1.6 nazivamo funkcija gustoće vjerojatnosti od X i kraće je označavamo sa f.

Definicija 1.1.11. *Ako red* $\sum_{\omega_k \in \Omega} X(\omega_k)\mathbb{P}(\{\omega_k\})$ *apsolutno konvergira, onda njegovu sumu zovemo matematičko očekivanje ili, kraće, očekivanje slučajne varijable X i označujemo sa*

$$\mathbb{E}(X) = \sum_{\omega_k \in \Omega} X(\omega_k)\mathbb{P}(\{\omega_k\}). \tag{1.7}$$

Jasno je da u slučaju konačnog skupa ω svaka slučajna varijabla ima očekivanje.

Definicija 1.1.12. *Neka je X slučajna varijabla i neka* $\mathbb{E}(X)$ *postoji. Varijanca od X definira se kao*

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2] \tag{1.8}$$

ako očekivanje u 1.8 postoji.

Neka slučajna varijabla X ima varijancu. Standardna devijacija σ_X od X nenegativan je kvadratni korijen iz varijance, tj.

$$\sigma_X = \sqrt{\mathrm{Var}X}. \tag{1.9}$$

Teorem 1.1.13. *Neka su X i Y slučajne varijable na* Ω *te neka postoje* $\mathbb{E}(X^2)$, $\mathbb{E}(Y^2)$. *Tada postoji i* $\mathbb{E}(X, Y)$ *te vrijedi*

$$|\mathbb{E}(X, Y)| \leq (\mathbb{E}(X^2)\mathbb{E}(Y^2))^{\frac{1}{2}}. \tag{1.10}$$

Definicija 1.1.14. *Neka je $X = (X_1, \ldots, X_n)$ n-dimenzionalni slučajni vektor na Ω i neka postoji $\mathbb{E}(X_i)^2$ za $i = 1, \ldots, n$). Iz 1.10 slijedi da postoje realni brojevi*

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))).$$

Za $i \neq j$ taj broj zovemo kovarijanca slučajnih varijabla X_i, X_j.

Teorem 1.1.15 (Zakon ukupne kovarijance)**.** *Neka su X, Y i Z slučajne varijable definirane na istom vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$ te neka je $\mathrm{Cov}(X, Y) < \infty$. Tada vrijedi*

$$\mathrm{Cov}(X, Y) = \mathbb{E}(Cov(X, Y|Z) + \mathrm{Cov}(E(X|Z, E(Y|Z)). \tag{1.11}$$

Definicija 1.1.16. *Neka su $X_1, \ldots, X_n$ slučajne varijable na vjerojatnosnom prostoru $(\Omega, \mathcal{F}, \mathbb{P})$. Kažemo da su $X_1, \ldots, X_n$ nezavisne slučajne varijable ako za proizvoljne $B_i \in \mathscr{B}$ ($i = 1, \ldots, n$) vrijedi*

$$\mathbb{P}\{X_1 \in B_1, \ldots, X_n \in B_n\} = \mathbb{P}\Big(\bigcap_{i=1}^{n}\{X_i \in B_i\}\Big) = \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\}. \tag{1.12}$$

Definicija 1.1.17. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira gotovo sigurno (g.s.) prema slučajnoj varijabli X ako je*

$$\mathbb{P}\Big\{\omega \in \Omega;\ X(\omega) = \lim_{n\to\infty} X_n(\omega)\Big\} = 1. \tag{1.13}$$

To označujemo (g.s.) $X_n \xrightarrow{g.s} X$, $(n \to \infty)$.

Definicija 1.1.18. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira po vjerojatnosti prema slučajnoj varijabli X ako za svaki $\varepsilon > 0$ vrijedi*

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| \geq \varepsilon\} = 0. \tag{1.14}$$

To označujemo kao $X_n \xrightarrow{\mathbb{P}} X$, $(n \to \infty)$.

Definicija 1.1.19. *Kažemo da niz $(X_n, n \in \mathbb{N})$ slučajnih varijabli konvergira po distribuciji prema slučajnoj varijabli X ako je*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \qquad za\ x \in C(F_X). \tag{1.15}$$

To označujemo kao $X_n \xrightarrow{\mathcal{D}} X$, $(n \to \infty)$.

Zakoni velikih brojeva

Teorem 1.1.20 (Slabi zakon velikih brojeva)**.** *Neka je* $(X_n, n \in \mathbb{N})$ *niz slučajnih varijabli takav da su za svaki* $n \in \mathbb{N}$ *varijable* $X_1, X_2, \ldots, X_n$ *nezavisne i neka postoji realan broj* $M > 0$ *takav da je* $\mathrm{Var}(X_k) \leq M, k \in \mathbb{N}$*. Stavimo* $S_n = \sum_{k=1}^{n} X_k, n \in \mathbb{N}$*. Tada za svako* $\varepsilon > 0$ *vrijedi*

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{S_n - \mathbb{E}S_n}{n}\right| \geq \varepsilon\right\} = 0, \tag{1.16}$$

tj.

$$\frac{S_n - \mathbb{E}S_n}{n} \xrightarrow{\mathbb{P}} 0 \quad za \quad n \to \infty. \tag{1.17}$$

Teorem 1.1.21 (Jaki zakon velikih brojeva)**.** *Neka je* $(X_n, n \in \mathbb{N})$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli. Tada niz* $\left(\frac{1}{n}\sum_{j=1}^{n} X_j, n \in \mathbb{N}\right)$ *konvergira (g.s.) ako i samo ako* $\mathbb{E}X_1$ *postoji i u tom slučaju je*

$$\frac{1}{n}\sum_{j=1}^{n} X_j \xrightarrow{g.s} \mathbb{E}X_1 \quad za \quad n \to \infty. \tag{1.18}$$

Centralni granični teorem

Teorem 1.1.22 (Levy)**.** *Neka je* $(X_n, n \in \mathbb{N})$ *niz nezavisnih, jednako distribuiranih slučajnih varijabli s očekivanjem* m *i varijancom* σ^2*,* $0 < \sigma^2 < \infty$ *i neka je* $S_n = \sum_{k=1}^{n} X_k$*,* $n \in \mathbb{N}$*. Tada vrijedi*

$$\frac{S_n - \mathbb{E}S_n}{\sigma\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1), \qquad za\ n \to \infty. \tag{1.19}$$

Neke poznate funkcije distribucije

Definicija 1.1.23. *Neka su* $m, \sigma \in \mathbb{R}$*, pri čemu je* $\sigma > 0$*. Neprekidna slučajna varijabla* X *ima normalnu distribuciju s parametrima* m *i* σ^2*, ako joj je gustoća* f *dana sa*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-m)^2}{2\sigma^2}}. \tag{1.20}$$

Oznaka koju koristimo jest $X \sim N(m, \sigma^2)$*.*

Definicija 1.1.24. *Neprekidna slučajna varijabla X ima eksponencijalnu distribuciju s parametrom λ ako joj je gustoća f dana sa*

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{1.21}$$

Oznaka koju koristimo jest $X \sim Exp(\lambda)$.

Definicija 1.1.25. *Neprekidna slučajna varijabla X ima uniformnu distribuciju na segmentu $[a, b]$ za $a, b \in \mathbb{R}$, ako joj je gustoća f dana sa*

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & x \notin [a, b]. \end{cases} \tag{1.22}$$

Oznaka koju koristimo jest $X \sim U(a, b)$.

Definicija 1.1.26. *Neka su $a, b \in \mathbb{R}$ i $a > 0$. Neprekidna slučajna varijabla X ima Cauchyjevu distribuciju s parametrima a i b ako joj je gustoća f dana sa*

$$f(x) = \frac{a}{\pi[a^2 + (x-b)^2]}, \qquad x \in \mathbb{R}. \tag{1.23}$$

Oznaka koju koristimo jest $X \sim C(a, b)$.

Definicija 1.1.27. *Neka je $\alpha > 0, \beta > 0$ i $\Gamma(X) = \int_0^\infty e^{-t} t^{x-1} dt, x > 0$, tj. Γ je gama-funkcija. Neprekidna slučajna varijabla X ima gama-distribuciju s parametrima α, β ako joj je gustoća f dana sa*

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{\frac{-x}{\beta}}, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{1.24}$$

Oznaka koju koristimo jest $X \sim \Gamma(\alpha, \beta)$.

Definicija 1.1.28. *Neka je $n \in \mathbb{N}$. Neprekidna slučajna varijabla X ima Studentovu t-distribuciju sa n stupnjeva slobode ako joj je gustoća f dana sa*

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{\frac{-(n+1)}{2}}, \qquad x \in \mathbb{R}. \tag{1.25}$$

Oznaka koju koristimo jest $X \sim t_n$.

Definicija 1.1.29. *Neka je $\alpha > 0, x_m > 0$. Neprekidna slučajna varijabla X ima Paretovu distribuciju ako joj je gustoća f dana sa*

$$f(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}}, & x \geq x_m, \\ 0, & x < x_m. \end{cases} \tag{1.26}$$

Oznaka koju koristimo jest $X \sim Pareto(\alpha, x_m)$.

Definicija 1.1.30. *Neka je $k > 0, \lambda > 0$. Neprekidna slučajna varijabla X ima Weibullovu distribuciju ako joj je gustoća f dana sa*

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{1.27}$$

Metoda najveće vjerodostojnosti

Većina pojmova iz ovog poglavlja preuzeta je iz [16].

Definicija 1.1.31. *Slučajni uzorak duljine n na statističkoj strukturi $(\Omega, \mathcal{F}, \mathbb{P})$ jest niz $X_1, X_2, \ldots, X_n$ slučajnih veličina na $(\Omega, \mathcal{F})$ takvih da su nezavisne i jednako distribuirane u odnosu na svaku vjerojatnost $\mathbb{P} \in \mathcal{P}$.*

Definicija 1.1.32. *Statistika na statističkoj strukturi $(\Omega, \mathcal{F}, \mathbb{P})$ jest svaka slučajna veličina koja je izmjeriva funkcija nekog slučajnog uzorka na toj statističkoj strukturi.*

Definicija 1.1.33. *Neka je X slučajni uzorak iz populacije s jednodimenzionalnim parametrom θ. $(1-\alpha)100\%$-pouzdan interval za θ je slučajni interval $[\hat{\theta}_1(X), \hat{\theta}_2(X)]$ takav da je*

$$\mathbb{P}(\hat{\theta}_1(X) \leq \theta \leq \hat{\theta}_2(X)) = 1 - \alpha. \tag{1.28}$$

Neka je $\mathbf{X} = (X_1, X_2, .., X_n)$ slučajni uzorak duljine $n, n \geq 1$ iz statističkog modela $\mathcal{P} = \{f(\cdot;\theta) : \theta \in \Theta\}$. Ako je $\mathbf{x} = (x_1, x_2, .., x_n)$ jedna realizacija od $\mathbf{X}$, tada je vjerodostojnost funkcija $L : \Theta \to \mathbb{R}$ definirana sa

$$L(\theta) = L(\theta|\mathbf{x}) := f_{\mathbf{X}}(\mathbf{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Theta. \tag{1.29}$$

Definicija 1.1.34. *Statistika $\hat{\theta} \equiv \hat{\theta}(\mathbf{X})$ je procjenitelj najveće (ili maksimalne) vjerodostojnosti ako vrijedi*

$$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta|\mathbf{X}) \tag{1.30}$$

Procjenu $\hat{\theta}(\mathbf{x}) = \hat{\theta}$ nazivamo procjena metodom maksimalne vjerodostojnosti ili kraće MLE (od eng.. *maximum likelihood estimate*).

Metoda najmanjih kvadrata

Neka su poznate vrijednosti zavisne varijable $Y = (y_1, y_2, \ldots, y_n)$ te nezavisne varijable $X = (x_1, x_2, \ldots, x_n)$. Prilagodbom jednodimenzionalnog modela linearne regresije, u obliku

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.31}$$

sumu kvadrata grešaka zapisujemo kao

$$S(\beta_1, \beta_0) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Rezultat metode najmanjih kvadrata je vektor koeficijenata $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ takav da je

$$S(\hat{\beta}) = \min S(\beta).$$

Postupak je analogan u višedimenzionalnom slučaju, kao što je i objašnjeno u [8].

1.2 Uvod u bootstrap metodu

Pojam bootstrap prvi put se spominje još u 19. stoljeću kao sinonim za nemoguć zadatak. Kao matematičko područje razvija se u 1979. objavom knjige Bradleya Efrona "Bootstrap Methods: Another Look at the Jackknife" koji je nadahnuće pronašao u Jackknife metodi.

Neka je $\mathbf{y} = (y_1, y_2, ..y_n)$ uzorak duljine n koji predstavlja realizaciju slučajnog uzorka $Y_1, Y_2, \ldots, Y_n$, pri čemu Y_i imaju funkciju gustoće koju označavamo s f, a funkciju distribucije s F. Na temelju danog uzorka želimo donijeti zaključak o karakteristici od interesa dane populacije, θ, koristeći statistiku T. Vrijednost statistike T u danom uzorku označavat ćemo s t. Pretpostavljamo da je statistika T definirana te je njezina realizacija t procjena za nepoznati parametar θ.

Problem od interesa je vjerojatnosna distribucija statistike T te razlikujemo dva pristupa rješavanju tog problema - parametarski i neparametarski. Kod parametarskog modela postoji određen matematički model definiran parametrima ψ koji potpuno određuju f. Samim time θ je komponenta ili funkcija od ψ. S druge strane, kada ne postoji matematički model koji bi određivao f, takav pristup nazivamo neparametarski te on koristi jedino pretpostavku da su varijable $Y_1, \ldots, Y_n$ nezavisne i jednako distribuirane.

Parametarski bootstrap

Pretpostavimo da je dan parametarski model definiran parametrom ψ koji opisuje distribuciju iz koje dolazi naš uzorak $\mathbf{Y} = (y_1, y_2, ..y_n)$. Označimo funkciju distribucije sa F_ψ, a pripadnu funkciju gustoće sa f_ψ. Koristeći metodu maksimalne vjerodostojnosti, iz uzorka procjenjujemo ψ sa $\hat{\psi}$ te na taj način dolazimo do tzv. *fitted modela*. Bootstrap uzorkovanje dalje provodimo oslanjajući se na $F_{\hat{\psi}}$. Ako su neke vrijednosti parametra ψ poznate, njih nije potrebno procjenjivati metodom maksimalne vjerodostojnosti.

U nastavku navodimo primjer procjene greške za očekivanje visine populacije dobivene bootstrap metodom.

Primjer 1.2.1. *Prema [21], prosječna visina za muškarce u Nizozemskoj je 183.8 cm, dok je standardna devijacija 7.1 cm. Dakle, funkcija distribucije je, zajedno sa svojim parametrima, poznata. Simuliran je uzorak 10 ljudi:*

1	2	3	4	5	6	7	8	9	10
179.3	185.1	177.8	195.1	186.1	177.9	187.2	189.0	187.8	181.6

Statistika koju promatramo je srednja vrijednost. Vrijednost statistike na uzorku je

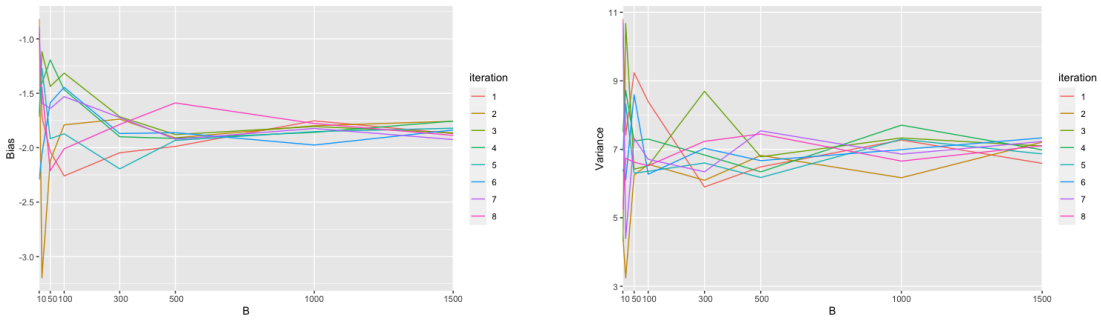
$$\bar{x} = 184.7386$$

Bootstrap metodom za neki B, simuliramo ukupno B različitih uzorka duljine $n = 10$ *iz normalne distribucije čiji su parametri poznati,* $N(183.8, 7.1^2)$*. Za svaki uzorak računamo statistiku od interesa* $T_1^*, T_2^*, .., T_B^*$*. Da bismo znali koliko je ta procjena dobra za određeni broj simulacija B, koristimo mjeru pristranosti i varijance.*

$$bias_B = \bar{T}^* - t \tag{1.32}$$

$$var_B = \frac{1}{B-1} \sum_{b=1}^{B} (T_b^* - \bar{T}^*)^2 \tag{1.33}$$

Provedeno je ukupno 8 iteracija za $B = 10, 50, 100, 300, 500, 1000, 1500$*. Rezultati su prikazani na slici 1.1*



Slika 1.1: Simulirane vrijednosti za pristranost i varijancu

Primjećujemo kako se bootstrap procjena varijance statistike smanjuje kako se poveća broj simuliranja B, no također vidimo da i dobivene vrijednosti var_B *i* $bias_B$ *konvergiraju egzaktnim vrijednostima normalne distribucije kada* $B \to \infty$*. Jedno od pitanja koje se ovdje nameće jest koji je "dovoljno dobar" broj uzorkovanja B? No o ovome će biti riječ nešto kasnije.*

Neparametarski bootstrap

Pretpostavimo da su $Y_1, .., Y_n$ nezavisne, jednako distribuirane slučajne varijable s nepoznatom funkcijom distribucije F. Kod neparametarskih modela, ključna je uloga empirijske funkcije distribucije koja dodjeljuje jednake vjerojatnosti svakoj vrijednosti slučajnog uzorka. Dakle, funkciju distribucije u ovom slučaju procjenjujemo koristeći empirijsku funkciju distribucije $\hat{F}$:

$$\hat{F}(x) = \frac{\sharp\{y_j \leq y\}}{n}.$$

Empirijska funkcija distribucije ima istu ulogu kao i funkcija distribucije u parametarskom modelu čiji su parametri procijenjeni iz danog uzorka.

Simuliranje podataka koristeći empirijsku funkciju distribucije prilično je jednostavno. Naime, budući da su iste vjerojatnosti dodijeljenje svim vrijednostima originalnog uzorka $\mathbf{y} = (y_1, y_2, ..y_n)$, novo simuliran uzorak $\hat{\mathbf{y}} = (\hat{y_1}, \hat{y_2}, ..\hat{y_n})$ sadržava iste vrijednosti kao i uzorak $\mathbf{y}$ s mogućim ponavljanjem.

U nastavku iznosimo primjer u kojem se proučava utjecaj veličine uzorka pri korištenju neparametarske bootstrap metode.

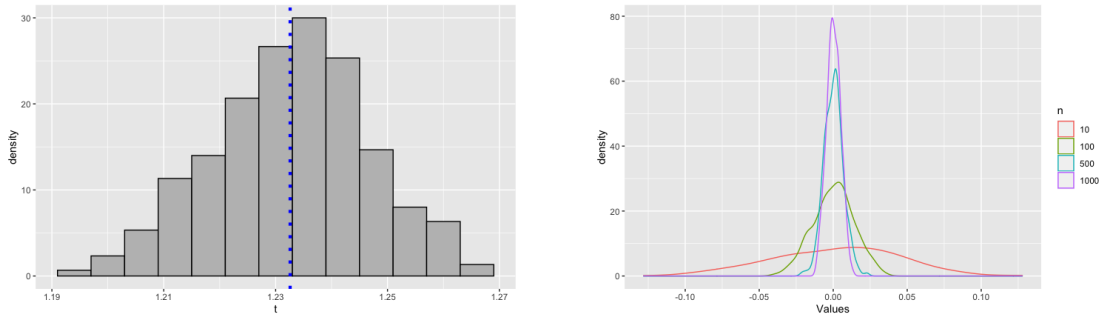
Primjer 1.2.2. *Koristimo transformirane podatke preuzete sa [1].*
Ukupan uzorak čini 55 692 *pacijenta, no zbog ilustracije uzimat ćemo prvih nekoliko vrijednosti. Proučavamo sljedeće varijable:*

$$X = \textit{obujam struka pacijenta, mjeren u cm}$$
$$Y = \textit{tjelesna masa pacijenta, mjerena u kg}$$

Jedan od ključnih indikatora rane dijagnoze dijabetesa tipa 2, kako je objašnjeno u [22], jest omjer obujma struka i tjelesne mase. Stoga je statistika od interesa:

$$T = \frac{E(X)}{E(Y)} \tag{1.34}$$

Uzorkovanjem $B = 500$ *puta iz uzorka za različite veličine* $n = 10, 100, 500, 1000$*, računamo statistiku od interesa. Rezultati su prikazi na slici 1.2.*



(a) Histogram vrijednosti **(b)** Procijenjene funkcije gustoće

Slika 1.2: Rezultati simulacije neparametarskom bootstrap metodom

Lijevi histogram prikazuje distribuciju dobivenih vrijednosti statistike 1.34, za $n = 100$*. Plavom isprekidanom linijom označena je vrijednost statistike na danom uzorku.*
Desno su prikazane procijenjene funkcije gustoće za $T^* - t$ *uzimajući u obzir različite veličine uzorka. Što je veći uzorak, vidljivo je da je prilagodba normalnom modelu bolja.*

Poglavlje 2

Jednostavna linearna regresija

2.1 Prilagodba modela

Neka je Y varijabla odaziva (zavisna varijabla), a X kovarijata (nezavisna varijabla). Pretpostavimo da je poznato n vrijednosti tih slučajnih varijabli te označimo njihove realizacije sa $y_1, \ldots, y_n$ i $x_1, x_2, \ldots, x_n$. Jednostavan model linearne regresije dan je sa

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \qquad j = 1, \ldots, n \tag{2.1}$$

pri čemu su ε_j nekorelirane, homeskedastične greške[1] s očekivanjem 0 i jednakom varijancom σ^2. Procjene metodom najmanjih kvadrata za β dane su kao

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})y_j}{SS_x}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.2}$$

pri čemu je $\bar{x} = n^{-1} \sum x_j$, a $SS_x = \sum_{j=1}^{n}(x_j - \bar{x})^2$.
Procjena greške varijance σ^2 je

$$s^2 = \frac{1}{n-2} \sum_{j=1}^{n} e_j^2, \tag{2.3}$$

pri čemu definiramo

$$e_j = y_j - \hat{\mu}_j \tag{2.4}$$

kao reziduale linearne regresije, a

$$\hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j \tag{2.5}$$

[1] Pod pojmom homoskedastičnost grešaka podrazumijeva se jednakost varijance.

kao procijenjene vrijednosti od y_j za dani x_j.
Osnovna svojstva procjenitelja $\hat{\beta}_0$, $\hat{\beta}_1$ su

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \qquad \text{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_x} \right) \tag{2.6}$$

i

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \qquad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{SS_x}. \tag{2.7}$$

Procjenitelji su normalno distribuirani i optimalni ukoliko su greške ε_j normalno distribuirane, a često približno normalno distribuirani za druge distribucije grešaka. Ipak, nisu otporni na veliku "nenormalnost" grešaka. Više o ovome može se pronaći u [6].

2.2 Pretpostavke linearne regresije

Prilagodba linearne regresije metodom najmanjih kvadrata daje zadovoljavajuće procjenitelje u slučaju kada su zadovoljene sve pretpostavke modela. Olako shvaćanje važnosti tih pretpostavka i njihovo zanemarivanje može dovesti do katastrofalnih rezultata. Nekonzistentnost procjenitelja, pogrešni intervali pouzdanosti, premale ili prevelike p vrijednosti - samo su neke od njih. U ovom poglavlju razrađene su četiri temeljne pretpostavke linearne regresije pod kojima metoda najmanjih kvadrata daje ispravne, konzistentne procjenitelje. Ovo poglavlje većinom je inspirirano [10] i [18].

- Linearni odnos između varijabla poticaja i odaziva
  Pretpostavljamo da je uvjetno očekivanje grešaka jednako 0 za bilo koje vrijednosti nezavisne varijable X. Ovo povlači to da veza nezavisne varijable X te populacijskog očekivanje varijable Y mora biti linearna. Ako postoji neki nelinearan odnos u modelu, predikcijske vrijednosti mogu biti potpuno promašene.

- Normalna distribuiranost grešaka
  Greške moraju dolaziti iz normalne distribucije. Kao što ćemo vidjeti u nastavku, velika odstupanja od normalnosti mogu prouzročiti brojne probleme pri procjeni vrijednosti i pouzdanih intervala. Normalnost grešaka možemo provjeriti crtanjem normalnog vjerojatnosnog grafa reziduala ili korištenjem statističkih testova (Lillieforsov test, Shapiro-Wilkov test...). Veliki uzorci koji pripadaju distribuciji lakog repa nisu problematični - u tom slučaju greške su potpuno kontrolirane i zbog snage centralnog graničnog teorema svi rezultati vrijede. Problemi su mogući kada greške imaju distribuciju teškog repa. Više o ovome može se naći u [14].

- Homoskedastičnost grešaka[2]
  Varijanca grešaka je konstantna za sve vrijednosti nezavisne varijable. Ova pretpostavka, kombinirana zajedno s prethodnom, povlači da greške dolaze iz normalne distribucije s fiksnom standardnom devijacijom. U višedimenzionalnoj regresiji nejednakost varijance grešaka može rezultirati preferiranjem određenog podskupa podataka pri prediktiranju novih vrijednosti. Heteroskedastičnost se najčešće pojavljuje kod vremenskih podataka gdje varijanca raste s vremenom. Jedan dobar primjer u praksi naveden je u [4]. Također, pretpostavka o zadovoljenosti homoskedastičnosti može se provjeriti crtanjem odnosa reziduala i prediktiranih vrijednosti (*residual-fit plot*).

- Nezavisnost grešaka
  Slučajne greške $\varepsilon_1, \ldots, \varepsilon_n$ moraju biti međusobno nezavisne. Zaključci o nezavisnosti grešaka najčešće se izvode koristeći grafički prikaz autokorelacijske funkcije. Zavisnost se često javlja u praksi - kod tzv. *cluster*[3] podataka i često kod longitudinalnih podataka. U tim slučajevima koriste se specijalni modeli koji uključuju koreliranost podataka - npr. ARIMA modeli.

Ako su zadovoljene gornje navedene pretpostavke - greške modela su homoskedastične, slučajne i normalno distribuirane ili ako na raspolaganju imamo puno podataka, tada možemo parametre modela procijeniti metodom najmanjih kvadrata i rezultati će biti dovoljni dobri.

U [7] se iznosi primjer prilagodbe modelu s heteroskedastičnim nenormalnim greškama i objašnjavaju mogućnosti koje nude metode ponovnog uzorkovanja u tom slučaju. Dakle, za heteroskedastične, nenormalne, slučajne greške s nejednakom varijancom rješenja metode najmanjih kvadrata nisu dovoljno zadovoljavajuća te alternativu tražimo u metodama ponovnog uzorkovanja kao što će biti i objašnjeno u nastavku.

2.3 Svojstva reziduala

Reziduali modela linearne regresije od velike su važnosti za ispitivanje značajnosti i prilagodbe modelu linearne regresije. Ponekad su upravo oni ti koji mogu "otkriti" koliko je dobar ili loš model koji smo prilagodili podacima.

Također, budući da daju procjene za slučajne greške ε_j, važnu primjenu pronalaze i u metodama ponovljenog uzorkovanja.

[2] U literaturi se često upotrebljava i termin homogenost grešaka.

[3] Podaci su uzorkovani iz jedne određene grupe, umjesto iz cijele populacije.

Pod pretpostavkom modela 2.1, reziduale možemo zapisati kao

$$e_j = \sum_{k=1}^{n} h_{j,k}\varepsilon_k, \tag{2.8}$$

pri čemu su

$$h_{j,k} = n^{-1}\delta_{j,k} + \frac{(x_j - \bar{x})(x_k - \bar{x})}{SS_x}. \tag{2.9}$$

Ovdje je $\delta_{j,k}$ Kronecker delta simbol:

$$\delta_{ij} = \begin{cases} 1, & \text{za } i = j, \\ 0, & \text{za } i \neq j. \end{cases}$$

Veličine $h_{j,j}$ zovemo poluge i označavamo ih kraće sa h_j. Iz 2.8 i linearnosti matematičkog očekivanja slijedi da

$$\mathbb{E}(e_j) = 0. \tag{2.10}$$

Ako veličine $h_{i,j}$ posložimo u matricu i označimo tu matricu sa H, ona se može izraziti još i kao $H = x(x^T x)^{-1}x^T$[4], pri čemu je x vektor opservacija $x = (x_1, \ldots, x_n)$, odakle slijedi

$$\begin{aligned} \text{var}(e_j) &= \text{var}(y - \hat{y}) \\ &= \text{var}((I - H)y) \\ &= (I - H)\text{var}(y)(I - H)^T \\ &= \sigma^2(I - H)^2. \end{aligned}$$

Dakle,

$$\text{var}(e_j) = \sigma^2(1 - h_j). \tag{2.11}$$

Vidimo da, ako su slučajne greške ε_j nezavisne i jednako distribuirane, ostaci e_j ne moraju biti jednako distribuirani. [5]
Najpogodnije rješenje ovog problema jest da transformiramo reziduale e_j u oblik u kojem

[4] U terminologiji pojmova vektorskih prostora H je ortogonalna projekcija na potprostor koji razapinju stupci matrice X.

[5] Za više intuicije i razumijevanja izraza varijance čitatelja se upućuje na [3].

oni imaju konstantnu varijancu. S tim ciljem, podijelit ćemo ih s njihovom standardnom devijacijom

$$t_j = \frac{e_j}{(\sigma^2(1-h_j)^{\frac{1}{2}}}. \tag{2.12}$$

Ovakve reziduale nazivamo studentizirani reziduali. Naziv dolazi iz činjenice da, pod pretpostavkom da greške dolaze iz normalne distribucije s očekivanjem 0 i varijancom σ^2, oni slijede studentovu distribuciju sa $(n-2)$ stupnjeva slobode, kao što je i pokazano u [19].

Također, postoje i razne druge transformacije reziduala s ciljem detekcije stršećih vrijednosti. Znatiželjne čitatelje upućujemo na poglavlje 10 u [17].

Modificirani reziduali su identični studentiziranim rezidualima, ali bez dijeljenje uzoračkom standardnom devijacijom. U mnogim primjerima u ovom radu koristit ćemo upravo njih

$$r_j = \frac{y_j - \bar{\mu}_j}{(1-h_j)^{\frac{1}{2}}}. \tag{2.13}$$

U praksi Q-Q dijagram[6] modificiranih reziduala veoma jasno će otkriti očite ekstremne vrijednosti i/ili nenormalnost slučajnih grešaka.

2.4 Alternativni pristupi linearnoj regresiji

Potreba za prilagodbom modela linearne regresije može se pojaviti i na druge načine kao što je opisano u [15]. Važnost tih rezultata je u tome što služe kao temelj za razvijanje bootstraping metoda.

Bivarijantna distribucija $F(X, Y)$

Parove podataka (x_j, y_j) promatramo kao da pripadaju bivarijantnoj distribuciji $F(X, Y)$. Pretpostavljamo da je uvjetno očekivanje od Y uz dani $X = x$, linearno po x.

Iz pretpostavke linearnosti, imamo

$$\mathbb{E}(Y|X) = \beta_1 X + \beta_0,$$

pri čemu su β_0, β_1 realni brojevi. Uzimanjem matematičkog očekivanja s obje strane dobijemo

$$\mathbb{E}(\mathbb{E}(Y|X)) = \beta_1 \mathbb{E}X + \beta_0 \implies \mu_Y = \beta_1 \mu_x + \beta_0.$$

[6]Grafički prikaz teorijskih kvantila neke distribucije u ovisnosti o uzoračkim kvantilima danog uzorka.

Iz definicije kovarijance slijedi

$$\mathrm{Cov}(X, Y) = \rho\sigma_X\sigma_Y. \tag{2.14}$$

S druge strane, iz 1.11 za $Z = X$ imamo

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}(\mathrm{Cov}(X, Y|X)) + \mathrm{Cov}(\mathbb{E}(X|X), \mathbb{E}(Y|X)) \\ &= 0 + \mathrm{Cov}(X, \beta_0 + \beta_1 X) \\ &= \beta_1\sigma_X^2. \end{aligned} \tag{2.15}$$

Izjednačavanjem 2.14 i 2.15 dobivamo

$$\beta_1\sigma_X^2 = \rho\sigma_X\sigma_Y \implies \beta_1 = \rho\frac{\sigma_X}{\sigma_Y}. \tag{2.16}$$

Konačno, uvrštavanjem svih gornjih rezultata, slijedi

$$\begin{aligned}\mathbb{E}(Y|X) &= \beta_1 X + \beta_0 = \beta_1 X + \mu_Y - \beta_1\mu_X \\ &= \mu_Y + \beta_1(X - \mu_x) \\ &= \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mu_X). \end{aligned} \tag{2.17}$$

Dakle, linearna regresija javlja se u smislu linearnosti uvjetnog očekivanja od Y,

$$\mathbb{E}(Y|X = x) = \mu_Y + \gamma(x - \mu_x), \qquad \gamma = \sigma_{xy}/\sigma_x^2, \tag{2.18}$$

pri čemu su $\mu_x = \mathbb{E}(X)$, $\mu_y = \mathbb{E}(Y)$, $\sigma_x^2 = \mathrm{var}(X)$ te $\sigma_{x,y} = \mathrm{cov}(X, Y)$.

Uvjetno očekivanje u 2.18 odgovara očekivanju definiranom u 2.1, ako stavimo

$$\beta_0 = \mu_y - \gamma\mu_x, \qquad \beta_1 = y.$$

Dakle, parametar $\beta = (\beta_0, \beta_1)^\intercal$ u ovom slučaju je funkcija prvog i drugog momenta funkcije distribucije $F(X, Y)$.

Niz distribucija F_x

Drugi način na koji možemo pristupiti problemu prilagodbe modela jest da za svaku točku x vrijednost varijable odaziva Y_x promatramo kao da pripada distribuciji $F_x(y)$, čije je očekivanje $\mu(x) = \beta_0 + \beta_1 x$, a varijanca $\sigma^2(x)$.

Očito je $\beta_0 = \mu(0)$ te je β_1 linearna kombinacija $\mu(x_1), \mu(x_2), \ldots, \mu(x_n)$, odnosno:

$$\beta_1 = \frac{\sum(x_j - \bar{x})\mu(x_j)}{SS_x}. \tag{2.19}$$

Za jednostavnu linearnu regresiju u kojoj su greške homoskedastične, ako označimo sa G distribuciju slučajnih grešaka s očekivanjem nula i varijancom σ[27], možemo koristiti

$$F_x(y) \equiv G\{y - \mu(x)\}. \tag{2.20}$$

Dakle, kako bismo mogli primijeniti ovaj pristup, potrebno je odrediti distribuciju F_x koristeći 2.20. To radimo s pomoću distribucije slučajnih grešaka te procijenjenih vrijednosti za x koje računamo koristeći parametre dobivene metodom najmanjih kvadrata. No, ostaje pitanje kako procijeniti distribuciju slučajnih grešaka jer su one nepoznate.

2.5 Uzorkovanje podataka

Uzorkovanje podataka linearne regresije je način korištenja bootstrap metode za procjenu greške procjenitelja linearne regresije.

Ako pretpostavimo da podaci za koje prilagođavamo model, dolaze iz bivarijantne distribucije $F(X, Y)$, kao što je objašnjeno u prvom dijelu 2.4, koeficijenti su tada funkcije od F definirani u 2.18.

U tom slučaju ne pretpostavlja se normalnost slučajnih grešaka ε_j, već samo njihova nezavisnost. Za bivarijantnu distribuciju $\hat{F}(X, Y)$ uzimamo empirijsku funkciju koja pridaje jednake vjerojatnosti svakom paru podataka.

U svakom koraku metode, nakon prilagodbe modela novo uzorkovanim podacima, koeficijenti se ponovo računaju metodom najmanjih kvadrata. Cijeli postupak opisan je algoritmom 1.

[7] Zbog pretpostavke o homoskedastičnosti grešaka vrijedi da je $\sigma(x) = \sigma$.

Algorithm 1 Uzorkovanje podataka za linearnu regresiju

1: **for** $r = 1, \ldots, R$ **do**
2: Odaberi $i_1^*, \ldots, i_n^*$ nasumično s ponavljanjem iz skupa $\{1, 2, 3, \ldots, n\}$.
3: **for** $j = 1, \ldots, n$ **do**
4: Neka je $x_j^* = x_{i_j^*}$.
5: Neka je $y_j^* = y_{i_j^*}$.
6: **end for**
7: Prilagodi linearni regresijski model metodom najmanjih kvadrata za podatke $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ te zapamti procijenjene koeficijente $\hat{\beta}_{0,r}^*$, $\hat{\beta}_{1,r}^*$, s_r^*.
8: **end for**

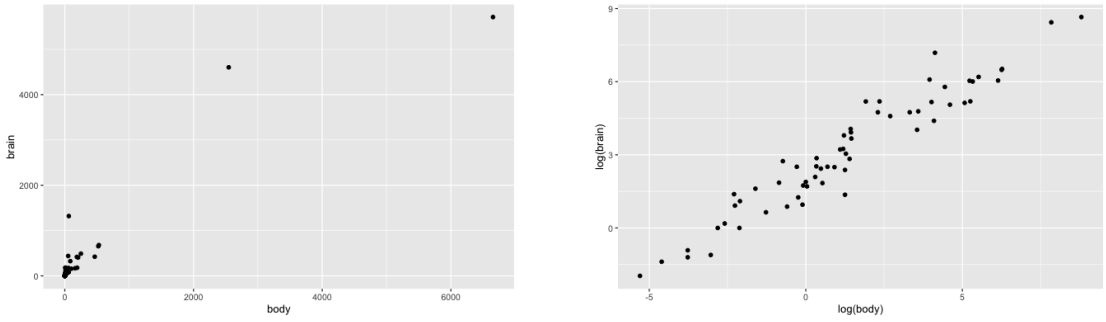
.

Primjer 2.5.1. *Opisanu metodu testiramo za set podataka Mammals*[8]

Dani su podaci za $n = 62$ *sisavca, prikazi na lijevoj strani slike 2.1. Grafički je teško očitati linearan odnos ovih podataka, stoga je očita potreba za transformacijom podataka, koju radimo logaritmiranjem. Varijable od interesa su:*

$$X = \log(\textit{tjelesna masa})$$
$$Y = \log(\textit{masa mozga})$$

Transformirani podaci prikazani su na desnoj strani slike 2.1.



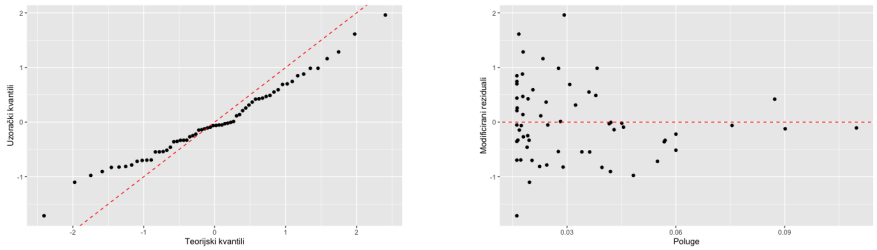
(a) Stvarni podaci **(b)** Logaritmirani podaci

Slika 2.1: Podaci za prilagodbu modela linearne regresije

Standardna analiza podataka sugerira blagu heteroskedastičnost, što se može naslutiti i iz grafičkog prikaza. Kao što je prikazano lijevo na slici 2.2, analiza reziduala i crtanje q-q dijagrama, potvrđujemo da greške slijede normalnu distribuciju jer se točke grupiraju oko pravca $y = x$. *Također, uočimo kako greške pripadaju distribuciji lakog repa.*

[8]Gotovi podaci su preuzeti iz R-ovog MASS paketa.

Analiza poluga prikazana je na desnoj strani slike 2.2. Poluge su mjere osjetljivosti procjenjenih vrijednosti $\hat{y}_i$ s obzirom na promjenu opažene vrijednosti y_i. Što je poluga veća, utjecaj pojedine opažene vrijednosti na predviđenu vrijednost je veći. Vidimo da u našem slučaju nemamo velikih poluga. Većina poluga ima vrijednost manju od 0.05.



(a) Usporedba s kvantilima normalne distribucije

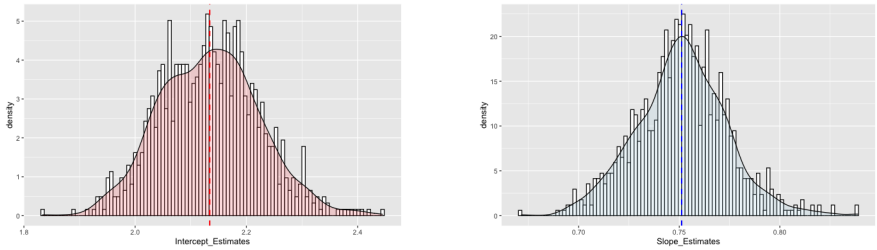
(b) Prikaz poluga

Slika 2.2: Analiza reziduala

Procjene parametara za prilagodbu modela linearne regresije dobivene metodom najmanjih kvadrata transformiranih podataka iznose:

$$\hat{\beta}_0 = 2.135$$
$$\hat{\beta}_1 = 0.752. \tag{2.21}$$

Uzorkovanjem podataka i korištenjem algoritma 1 dobiveni su sljedeći rezultati za $R = 1000$*:*[9]



(a) Koeficijent odsječka

(b) Koeficijent smjera

Slika 2.3: Histogram bootstrap procjenitelja

[9]Za svaku simulaciju koristi se sjeme 1234.

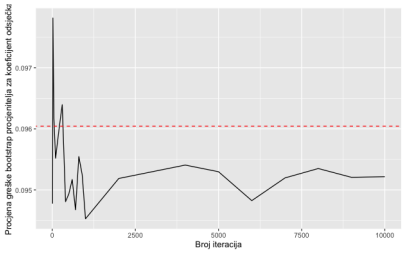
Uočava se normalna distribucija procjenitelja, iako nešto slabija nego kod prve metode. To pokazuje i izračun 5%-tnog i 95%-tnog kvantila studentiziranih vrijednosti koji malo više odudara od normalnih kvatila ± 1.645*:*

$$z^*_{(51)} = -1.584274$$
$$z^*_{(950)} = 1.696133.$$

Također, u nastavku su prikazani rezultati dobiveni za različit broj iteracija. Srednja standardna greška se i dalje stabilizira kako broj iteracija raste. Kod koeficijenta smjera ona je veoma blizu teorijskoj standardnoj greški.

Broj iteracija	Boostrap procjena	Procjena greške
10	2.126913	0.09477578
50	2.147071	0.09637237
100	2.144384	0.09541812
500	2.134127	0.0961766
1000	2.136078	0.09489997
5000	2.135607	0.09504343
10000	2.137028	0.09534211

(a) Rezultati

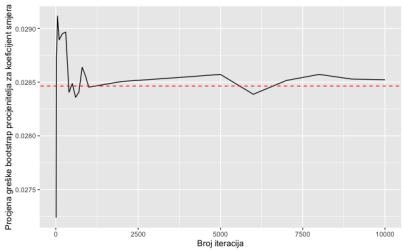


(b) Standardne greške

Slika 2.4: Uzorkovanje podataka za koeficijent odsječka

Broj iteracija	Boostrap procjena	Procjena greške
10	0.7512889	0.02723744
50	0.7570667	0.02900617
100	0.7494114	0.02876875
500	0.7509109	0.02898799
1000	0.7499268	0.02838711
5000	0.7515415	0.0285074
10000	0.751482	0.02855627

(a) Rezultati



(b) Standardne greške

Slika 2.5: Uzorkovanje podataka za koeficijent smjera

2.6 Uzorkovanje grešaka

Kako bismo iskoristili nešto modificiraniju metodu ponovljenog uzorkovanja za linearnu regresiju, identificirat ćemo model od interesa na malo drugačiji način.

Ako su greške u 2.1 uistinu homoskedastične, tada one dolaze iz točno jedne, određene distribucije. Uzmemo li da su x_j fiksni, primjenjiv je drugi alternativni pristup definiran u 2.4, pri čemu je G zajednička distribucija grešaka. Na model F gledamo kao na niz distribucija F_x za svaki $x_1, .., x_n$ definiranih u 2.20. U praksi za korištenje *resampling* metoda pripadne funkcije distribucije $\hat{F}_x$ procjenjujemo koristeći regresijske procjene [10] za $\mu(x_j)$, a procjenu za G dobivamo iz reziduala.

Kod parametarskog uzorkovanja pretpostavljamo da nam je poznata distribucija iz koje dolaze greške. Na primjer, pod pretpostavkom normalnosti, jedan od mogućih kandidata je $N(0, s^2)$, pri čemu je s kao u 2.3, no naravno da su u tom slučaju dostupni i teorijski rezultati, pa metoda uzorkovanja nije zanimljiva.

Za neparametarski pristup, koji nam je više od interesa, potrebna je generalizacija empirijske funkcije distribucije. Slučajne greške ε_j su nepoznate te samim time korištenje njihove empirijske funkcije distribucije nije moguće. Ideja je, budući da umjesto nepoznatih grešaka ε_j na raspolaganju imamo njihove procjene - reziduale e_j, iskoristiti njihovu empirijsku funkciju distribucije kao procjenu za G.

Ipak, za potpunu konzistentnost preporuča se korištenje modificiranih reziduala r_j, definiranih u 2.13, zbog podudarnosti njihove varijance s varijancom slučajnih grešaka,

$$\mathrm{Var}(r_j) = \mathrm{Var}\Big(\frac{y_j - \bar{\mu}_j}{\sqrt{1-h_j}}\Big) = \frac{1}{1-h_j}\sigma^2(1-h_j) = \sigma^2,$$

pri čemu smo koristili izraz za varijance reziduala izračunat u 2.11.

Također, iz pretpostavke da G ima očekivanje 0, kao konačnu procjenu za G koristimo empirijsku funkciju distribucije od $r_j - \bar{r}$, pri čemu je $\bar{r}$ prosjek od r_j. Tako definirani modificirani reziduali očito imaju očekivanje 0 te njihovu emprijsku funkciju distribucije označavamo sa $\hat{G}$, naglašavajući da je ona procjena za G.

[10] Pod pojmom regresijska procjena mislimo na vrijednosti $\hat{\mu}(x_j)$ definirane u 2.5

Za korištenje metoda ponovnog uzorkovanja model koristi isti dizajn kao i prije, do na sitnu modifikaciju. Dakle, i dalje koristimo $x_j{}^* \equiv x_j$, dok za uvjetnu distribuciju od $Y_j{}^*$ za dani $x_j{}^*$ koristimo procjenu od 2.1, odnosno,

$$y_j^* = \hat{\mu}_j + \varepsilon_j^* \qquad j = 1, 2, .., n, \tag{2.22}$$

pri čemu je $\hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j^*$ te ε_j^* nasumično uzorkovan iz funkcije distribucije $\hat{G}$. Cijeli postupak opisan je algoritmom 2.

Algorithm 2 Uzorkovanje grešaka za linearnu regresiju

1: **for** $r = 1, \ldots, R$ **do**
2: **for** $j = 1, \ldots, n$ **do**
3: Neka je $x_j^* = x_j$.
4: Nasumično izaberi ε_j^* iz $r_1 - \bar{r}, \ldots, r_n - \bar{r}$.
5: Neka je $y_j^* = \hat{\beta}_0 + \hat{\beta}_1 x_j^* + \varepsilon_j^*$.
6: **end for**
7: Prilagodi linearni regresijski model metodom najmanjih kvadrata za podatke $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ te zapamti procjene za $\hat{\beta}_{0,r}^*$, $\hat{\beta}_{1,r}^*$, s_r^*.
8: **end for**

Nakon dobivenih R bootstrap procjena iz gornjeg algoritma bootstrap procjenitelj dobiva se koristeći zakon velikih brojeva,

$$\hat{\beta}_0^* = \frac{\hat{\beta}_{0,1}^* + \hat{\beta}_{0,2}^* + \cdots + \hat{\beta}_{0,R}^*}{R}, \tag{2.23}$$

$$\hat{\beta}_1^* = \frac{\hat{\beta}_{1,1}^* + \hat{\beta}_{1,2}^* + \cdots + \hat{\beta}_{1,R}^*}{R}. \tag{2.24}$$

Pokažimo još kako se očekivanje i varijanca[11] procjenitelja koeficijenta odsječka $\hat{\beta}_1$ ponaša u skladu s teorijskim rezultatima. Bootstrap procjenitelj koeficijenta odsječka može se zapisati kao

$$\hat{\beta}_1^* = \frac{\sum(x_j - \bar{x})y_j^*}{\sum(x_j - \bar{x})^2} = \hat{\beta}_1 + \frac{\sum(x_j - \bar{x})\varepsilon_j^*}{SS_x}. \tag{2.25}$$

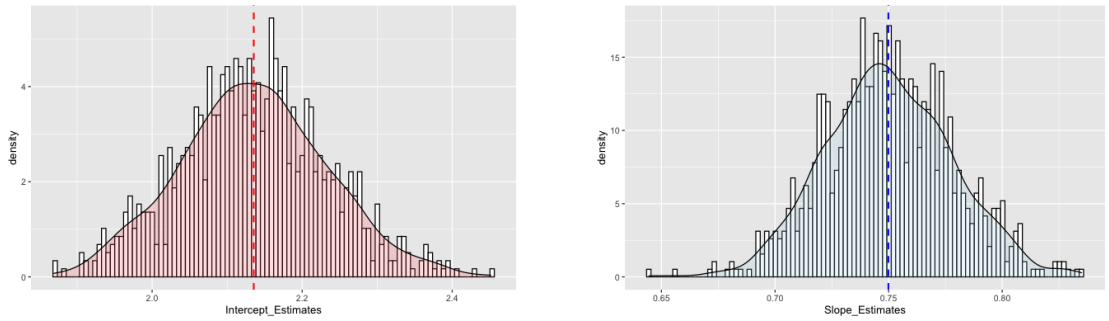
[11]Ovdje se promatra očekivanje i varijanca s obzirom na distribuciju uzorka, što je naglašeno oznakom zvjezdice.

Budući da je $\mathbb{E}^*(\varepsilon_j^*) = \dfrac{\sum(r_j - \bar{r})}{n} = 0$, slijedi da je $\mathbb{E}^*(\beta_1^*) = \hat{\beta}_1$.
Također, iz $\text{var}^*(\varepsilon_j^*) = \dfrac{\sum(r_j - \bar{r})^2}{n}$ za svaki $j = 1, \ldots, n$ slijedi

$$\text{var}^*(\beta_1^*) = \frac{\sum(x_j - \bar{x})^2 \text{var}(\varepsilon_j^*)}{SS_x^2} = n^{-1} \sum (r_j - \bar{r})^2 / SS_X. \tag{2.26}$$

Budući da je $n^{-1} \sum(r_j - \bar{r})^2 \doteq (n-2)^{-1} \sum e_j^2 = s^2$, slijedi da je 2.26 jednaka $\frac{s^2}{SS_x}$, što je upravo pokazano u 2.7.[12]

Primjer 2.6.1. *Na istom setu podataka kao i u primjeru 2.5.1, testirana je metoda uzorkovanja grešaka za procjenu koeficijenta. Korišten je algoritam 2 i programski kod naveden u 5. Simulacije su provedene za $R = 1000$ iteracija te su dobiveni sljedeći rezultati:*



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 2.6: Histogram bootstrap procjenitelja

Isprekidanom linijom prikazana je srednja vrijednost dobivenih procjenitelja. Lako se uočava da bootstrap procjenitelji zadovoljavajuće prate normalnu distribuciju.

Računanjem 5%-tnog i 95%-tnog kvantila studentiziranih procjenitelja dobivene su sljedeće vrijednosti:

$$z^*_{(51)} = -1.679263$$
$$z^*_{(950)} = 1.674070,$$

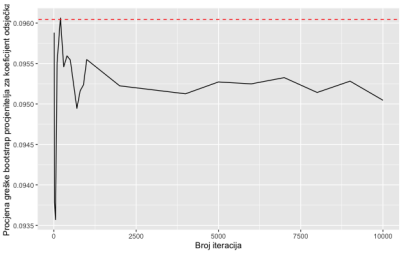
što je približno jednako normalnim kvantilima, koji iznose ± 1.645.
Dakle, očekivano, za veći broj podataka i "lijepe" greške rezultati uzorkovanja se podudaraju s teorijskim metodama.

[12] Oznaka $\doteq$ korištena je u smislu aproksimativne jednakosti, odnosno, jednakost se postiže kako $n \to \infty$.

Metoda je isprobana i za različit broj iteracija te su u nastavku navedene dobivene vrijednosti. Crvenom isprekidanom linijom označena je teorijska vrijednost navedena u 2.21. Uočimo kako se srednja standardna greška stabilizira za veći broj iteracija, što je posebno vidljivo u slučaju koeficijenta smjera.

Broj iteracija	Boostrap procjena	Procjena greške
10	2.151778	0.09588173
50	2.161196	0.09394577
100	2.126542	0.09482071
500	2.130905	0.09583942
1000	2.136133	0.09549899
5000	2.134965	0.09524924
10000	2.133488	0.09520729

(a) Rezultati

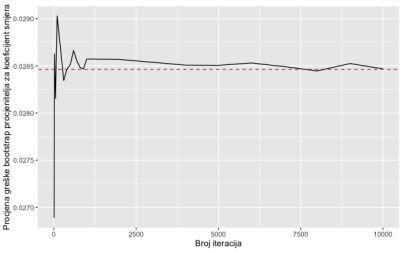


(b) Standardne greške

Slika 2.7: Uzorkovanje grešaka za koeficijent odsječka

Broj iteracija	Boostrap procjena	Procjena greške
10	0.7412849	0.02689087
50	0.7437914	0.02863644
100	0.7525952	0.02871952
500	0.7496104	0.02856015
1000	0.750452	0.02850296
5000	0.7515415	0.02851813
10000	0.7521315	0.02852106

(a) Rezultati



(b) Standardne greške

Slika 2.8: Uzorkovanje grešaka za koeficijent smjera

Usporedba metoda

Budući da imamo dva različita pristupa uzorkovanja za linearnu regresiju, pitanje koje se nameće samo od sebe jest: Koja bootstrap metoda je bolja? Postoje važne razlike između njih, kao što je to istaknuto u [9]. Svakako odgovor ovisi o tome koliko su ”jako” naše pretpostavke zadovoljene - koliko vjerujemo svom modelu?

Prva važna razlika jest da, ako uzorkujemo podatke, ne pretpostavljamo homoskedastičnost varijance - štoviše, čak ne pretpostavljamo ni da je uvjetno očekivanje od Y, uz dani $X = x$, linearno. U tom pogledu pruža se prilika za otpornost modela na heteroskedastičnost grešaka, no ostajemo bez efikasnosti rezultata ako se ispostavi da je varijanca

konstantna. Uzorkovanje podataka najčešće se preporuča kada su pretpostavke modela slabe i ne dovoljno pouzdane.

Druga važna razlika jest da uzorkovanjem podataka, budući da u svakoj iteraciji odabiramo slučajne vrijednosti za $x_1^*, \ldots, x_n^*$, svaki uzorak sadržava drugačije informacije za danu populaciju. Što znači da će varijabilnost u uzorkovanju $x_1^*, \ldots, x_n^*$ uzrokovati varijabilnost u informacijama dane populacije. No za velik broj podataka ovaj nedostatak najčešće neće biti od važnosti.

Kod uzorkovanja grešaka u algoritmu ostavljamo iste vrijednosti varijable X, što automatski smanjuje varijabilnost. Pretpostavke koje moramo zadovoljiti su "jače" - linearnost i homoskedastičnost, no dobiveni intervali pouzdanosti su uži nego kod uzorkovanja podataka. Detaljnija usporedba ovih metoda istaknuta je u [2].

Napomenimo još da u slučaju koreliranosti grešaka nijedna od ovih dviju metoda neće dati dobre rezultate.

U istaknutim primjerima vidljivo je da metoda uzorkovanja grešaka daje nešto bolje rezultate. Za najveći broj iteracija standardna greška je manja u metodi uzorkovanja grešaka tek u petoj decimali. Bootstrap procjenitelji koeficijenata su također "normalniji" u metodi uzorkovanja grešaka.

2.7 Nekonstanta varijanca - uzorkovanje grešaka različitih težina

U mnogim primjenama prilagodbe linearnog modela ispostavlja se da su greške heteroskedastične - nemaju konstantnu varijancu. Ako uspijemo tu varijabilnost objasniti i modelirati, bootstrap metode ostaju i dalje primjenjive prilikom procjene.
Podrazumijeva se da na početku koristimo metodu najmanjih kvadrata, identično kao i u poglavlju 2.1.

Poznata funkcija varijance

Pretpostavimo da je u 2.1, za slučajnu grešku ε_j u točki $x = x_j$, poznata varijanca $\sigma_j^2 = \kappa V(x_j)$ ili $\sigma_j^2 = \kappa V(\mu_j)$, pri čemu je $V(\cdot)$ poznata funkcija. Skalar κ moguće je procijeniti, no to nam nije od interesa u ovom trenutku.

Ideja kojom se želimo riješiti heteroskedastičnosti je vrlo jednostavna. Modificirane reziduale 2.13 dodatno skalirati vrijednostima poznate funkcije varijance kako bismo postigli barem približnu homoskedastičnost. Novi reziduali su:

$$r_j = \frac{y_j - \bar{\mu}_j}{V(x_j)(1-h_j)^{\frac{1}{2}}} \qquad \text{ili} \qquad r_j = \frac{y_j - \bar{\mu}_j}{V(\hat{\mu}_j)(1-h_j)^{\frac{1}{2}}}. \tag{2.27}$$

Nakon oduzimanja srednje vrijednosti $\bar{r}$ empirijska funkcija ovako definiranih reziduala poslužit će kao funkcija distribucije G skaliranih, homoskedastičnih slučajnih grešaka δ_j u modelu

$$Y_j = \beta_0 + \beta_1 x_j + V_j^{1/2}\delta_j, \tag{2.28}$$

pri čemu je $V_j = V(x_j)$ ili $V = V(\mu_j)$. Algoritam 2 modificiramo na sljedeći način:

Algorithm 3 Uzorkovanje grešaka sa nejednakom varijancom za linearnu regresiju

1: **for** $r = 1, \ldots, R$ **do**
2: **for** $j = 1, \ldots, n$ **do**
3: Neka je $x_j^* = x_j$.
4: Nasumično izaberi δ_j^* iz $r_1 - \bar{r}, \ldots, r_n - \bar{r}$.
5: Neka je $y_j^* = Y_j = \beta_0 + \beta_1 x_j + V_j^{1/2}\delta_j$, pri čemu je $V_j = V(x_j)$ ili $V = V(\mu_j)$, ovisno o zadanim početnim podacima.
6: **end for**
7: Prilagodi linearni regresijski model metodom najmanjih kvadrata za podatke $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ te zapamti procjene za $\hat{\beta}_{0,r}^*$, $\hat{\beta}_{1,r}^*$, s_r^*.
8: **end for**

Divlji bootstrap

Što ako je funkcija varijance slučajnih grešaka $V(\cdot)$ nepoznata? Ako postoji obrazac heteroskedastičnosti kod velikog skupa podataka, postoje metode kojima se njezine vrijednosti mogu modelirati koristeći vrijednosti reziduala. Na primjer, crtajući graf apsolutnih vrijednosti modificiranih reziduala r_j i procijenjenih vrijednosti $\hat{\mu}_j$. Ovaj pristup dobro prolazi ako u pozadini postoji jasna monotona veza između varijance sa x ili μ.

No ako nije jasno na koji se način ostvaruje heteroskedastičnost, rješenje ponovno nude metode ponovljenog uzorkovanja pokušavajući dati lokalnu procjenu varijance greške.

Takav "maksimalno" lokaliziran pristup naziva se divlji bootstrap te procjenjuje varijancu iz svakog reziduala individualno. Koristi se isti algoritam kao i prije, algoritam 2, jedino što j-tu grešku uzorkujemo iz specifične distribucije, raspodijeljenje u dvije točke:

$$\mathbb{P}\{\varepsilon_j^* = e_j(1 - \sqrt{5})/2\} = \frac{\sqrt{5}+1}{2\sqrt{5}}$$
$$\mathbb{P}\{\varepsilon_j^* = e_j(1 + \sqrt{5})/2\} = 1 - \frac{\sqrt{5}+1}{2\sqrt{5}}$$

pri čemu su $e_j = y_j - \hat{\mu}_j$ reziduali. Prva tri momenta ovako definirane distribucije su 0, e_j^2, e_j^3.
Rezultati su zadovoljavajući, no tek se u ovom stoljeću, pojavila ideja za novu, jednostavniju i efikasniju funkciju distribucije:[13]

$$\mathbb{P}\{\varepsilon_j^* = 1\} = \frac{1}{2}$$
$$\mathbb{P}\{\varepsilon_j^* = -1\} = \frac{1}{2}. \tag{2.29}$$

O tome koliko je ovako definirana funkcija distribucije važna za divlji boostrap kod heteroskedastičnog modela, pružajući bolje rezultate čak i od algoritma uzorkovanja podataka, govori rad [11].

[13] U literaturi se često koristi i naziv Rademacher distribucija.

Poglavlje 3

Višedimenzionalna linearna regresija

U ovom poglavlju proširujemo model jednostavne linearne regresije definiran u 2. Promatramo više kovarijata:

$$Y_j = \beta_0 x_{j0} + \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \varepsilon_j, \qquad j = 1, \ldots, n, \tag{3.1}$$

gdje je $x_{j0} \equiv 1$ za koeficijent odsječka različit od nule. U vektorskoj formi model glasi:

$$Y_j = x_j^T \beta + \varepsilon_j, \tag{3.2}$$

pri čemu je $x_j^T = (x_{j0}, x_{j1}, \ldots, x_{jp})$.

Matrična reprezentacija modela u kojoj označavamo sve vrijednosti varijable odaziva kao $Y^T = (Y_1, Y_2, \ldots, Y_n)$, glasi

$$Y = X\beta + \varepsilon \tag{3.3}$$

pri čemu je $X^T = (x_1, x_2, \ldots, x_n)$ te $\varepsilon^T = (\varepsilon_1, \ldots, \varepsilon_n)$.

Kao i prije, pretpostavljamo da su varijable Y_j nezavisne. U puno pogleda bootstrap analiza višedimenzionalnog linearnog modela jednostavno je proširenje bootstrap analize jednostavnog modela i opisanih algoritama u 2.

Ipak, neke situacije zahtjevaju dodatnu pozornost:

- procjena točnosti predikcije prilagođenog modela
- velika vrijednost parametra p u odnosu na n - relativno puno varijabli odaziva, no mala duljina uzorka
- selekcija "najboljeg" modela - biranje podskupa kovarijati koje najbolje opisuju varijablu odaziva.

3.1 Bootstrap uzorkovanje za metodu najmanjih kvadrata

Metoda najmanjih kvadrata daje procjene za vektor koeficijenata

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{3.4}$$

Procijenjene vrijednosti su $\hat{\mu} = Hy$, pri čemu je H matrica[1] koja se računa kao

$$H = X(X^T X)^{-1} X^T. \tag{3.5}$$

Dijagonalne elemente h_{jj} kraće označavamo kao h_j i zovemo poluge. Reziduale računamo kao

$$\epsilon = (I - H)y. \tag{3.6}$$

Pod pretpostavkom homoskedastičnosti slučajnih grešaka, procjena varijance za $\hat{\beta}$ je

$$\text{var}(\hat{\beta}) = s^2 (X^T X)^{-1}, \tag{3.7}$$

pri čemu je s^2 jednak $(n - p - 1)^{-1} e^T e$.
Aproksimacija varijance se može poboljšati kao i prije koristeći modificirane reziduale

$$r_j = \frac{e_j}{(1 - h_j)^{\frac{1}{2}}}. \tag{3.8}$$

Problem velikog broja kovarijata u odnosu na veličinu uzorka

Određene poteškoće javljaju se u obje bootstrap metode - metodi uzorkovanja grešaka i metodi uzorkovanja podataka u slučajevima kada je vrijednost parametra p velika u odnosu na n - veličinu uzorka. U nastavku je izložen ekstremni primjer takve situacije preuzet iz [15].

Primjer 3.1.1. *Neka je zadano ukupno m različith uzorka duljine n = 2. Matrica dizajna ima p = m stupaca te n = 2m redova, pri čemu je* $x_{2i-1,i} = x_{2i,i} = 1$, *inače* $x_{j,i} = 0$, *odnosno*

$$X = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}. \tag{3.9}$$

[1] Matrica H se često u literaturi naziva *hat* projektor.

Procjene koeficijenta su:

$$\hat{\beta}_i = \frac{1}{2}(y_{2i} + y_{2i-1}), \quad i = 1, \ldots, p. \tag{3.10}$$

Procjene reziduala iznose:

$$e_j = (-1)^j \frac{1}{2}(y_{2i} - y_{2i-1}), \quad h_j \equiv \frac{1}{2}, \quad j = 2i-1, 2i, \quad i = 1, \ldots, p. \tag{3.11}$$

Empirijska funkcija distribucije reziduala, čak i onih modificiranih, može jako odudarati od stvarne funkcije distribucije grešaka. Metoda uzorkovanja podataka će dati zadovoljavajuće procjene ako su slučajne greške homoskedastične. Metoda uzorkovanja podataka neće funkcionirati u slučaju kada u uzorku nedostaju y_{2+i}, y_{2i+2} zato što se tada koeficijent iz 3.10 ne može izračunati. Šanse da se to dogodi su 48% za $m = 5$, dok se za $m = 20$ povećavaju na 96%. Ovo se može popraviti tako da se u metodi maknu svi bootstrap uzorci u kojima je $f^*_{2i-1} + f^*_{2i} = 0$, za svaki i.

Zanimljivo je da se poteškoće uočene u gornjem primjeru mogu uočiti i u općenitim situacijama - za kombinacije $c^T\beta$, pri čemu c leži u potprostoru koji je razapet svojstvenim vektorima koji pripadaju malim svojstvenim vrijednostima matrice X^TX. Metoda uzorkovanja grešaka će dati zadovoljavajuće rezultate za procjenu standardne greške, no za računanje pouzdanih intervala za koeficijente ili predikciju bootstrap distribucija neće biti ni približno normalna. Što se tiče metode uzorkovanje podataka, ako je matrica $X^{*T}X^*$ blizu singularne, bootstrap procjene standardnih grešaka mogu biti lažno uvećane kao što je pokazano u primjeru 3.1.2.

Jedno od mogućih rješenja može biti da odbacimo sve bootstrap uzorke u kojima je najmanja svojstvena vrijednosti od $X^{*T}X^*$ manja od najmanje svojstvene vrijednosti X^TX. Alternativno rješenje jest da se gleda samo polovica uzorka - oni koji se grupiraju oko najmanje svojstvene vrijednosti matrice X.

Sve ove tvrdnje potkrijepljene su primjerom u nastavku.

Primjer 3.1.2. *Podaci izloženi u 3.1 poznati su u literaturi kao klasičan primjer visoko koreliranih podataka (koeficijenti koreliranosti dani su matricom korelacije u 3.2). Svaka od ukupno četiri kovarijate predstavlja postotak jedne od sastavnice betona.*

Najmanja svojstvena vrijednost od X^TX iznosi

$$\lambda = 0.001218 \tag{3.12}$$

te pripada svojstvenom vektoru

$$v = (0.99, -0.01, -0.01, -0.01, -0.01). \tag{3.13}$$

	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	1	68	8	12	109.4

Tablica 3.1: Podaci o cementu

	x_1	x_2	x_3	x_4	x_5
x_1	1.0000000	-0.9678114	-0.9977521	-0.9768711	-0.9982531
x_2	-0.9678114	1.0000000	0.9510277	0.9860529	0.9568429
x_3	-0.9977521	0.9510277	1.0000000	0.9623849	0.9979076
x_4	-0.9768711	0.9860529	0.9623849	1.0000000	0.9658978
x_5	-0.9982531	0.9568429	0.9979076	0.9658978	1.0000000

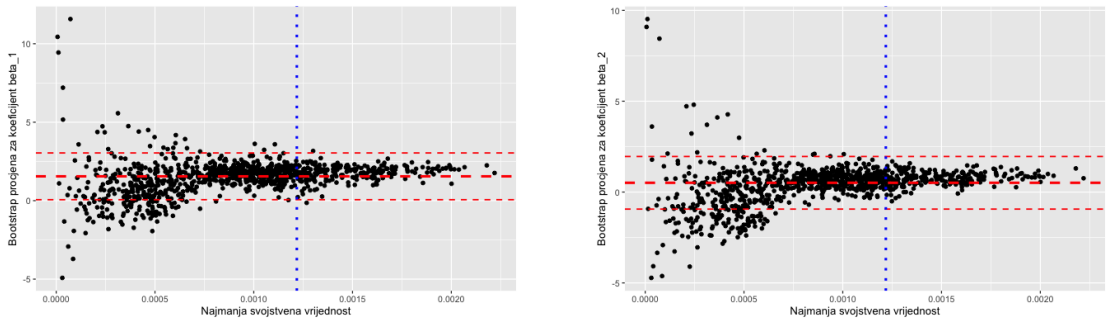
Tablica 3.2: Matrica korelacije

Teorijske i bootstrap standardne greške procijenjenih koeficijenta dane su u tablici 3.3. U prvom redu prikazane su teorijske vrijednosti. U drugom redu prikazani su rezultati dobiveni uzorkovanjem grešaka koji se više-manje slažu s teorijskim, kao što smo i očekivali. Uzorkovanje podataka daje veće standardne greške nego inače. Uzrok tomu može se naći u objašnjenju slika 3.1, 3.2.

Na slikama je prikazan odnos bootstrap procjene pojedinih koeficijenata te najmanje svojstvene vrijednosti matrice $X^{T*}X^*$. Vertikalnom plavom linijom označena je vrijednost najmanje svojstvene teorijske vrijednosti 3.12. Crvenom horizontalnom isprekidanom linijom označena je teorijska procjena koeficijenata te su označene linije pouzdanog intervala ± dvije standardne greške. Na svakom grafu uočava se isti obrazac. Varijabilnost u procjeni bilo kojeg od koeficijenata povećava se smanjenjem svojstvene vrijednosti. To sugerira ideju "izostavljanja" najmanjih svojstvenih vrijednosti te promatranje "novog" bootstrap uzorka. Rezultati takvog pristupa prikazani su u zadnjem redu tablice 3.3. Odmah vidimo veće poklapanje s teorijskim rezultatima.

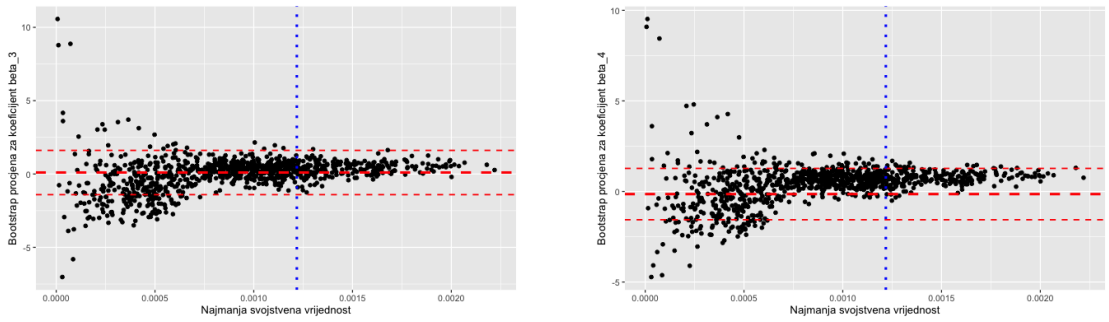
	β_0	β_1	β_2	β_3	β_4
Teorijska procjena	70.07	0.74	0.728	0.75	0.71
Uzorkovanje grešaka za R = 999	68.20	0.72	0.70	0.72	0.69
Uzorkovanje podataka za R = 999	123.17	1.33	1.28	1.30	1.27
Uzorkovanje podataka, srednjih 500	69.48	0.76	0.72	0.76	0.70
Uzorkovanje podataka, najvećih 800	62.86	0.69	0.65	0.68	0.64

Tablica 3.3: Standardne greške koeficijenata



(a) Koeficijent β_1 **(b)** Koeficijent β_2

Slika 3.1: Usporedba bootstrap procjenitelja i najmanjih svojstvenih vrijednosti.



(a) Koeficijent β_3 **(b)** Koeficijent β_4

Slika 3.2: Usporedba bootstrap procjenitelja i najmanjih svojstvenih vrijednosti.

3.2 Predikcija

Nakon prilagodbe modela linearne regresije postavlja se pitanje predikcije novih vrijednosti. Konkretno, zanima nas procijenjena vrijednost za Y_+ ako je vrijednost ulazne kovarijate x_+. Voljeli bismo opisati prediktiranu vrijednost pouzdanim intervalom. Dok se za procjenu $x_+^T\beta$ mogu koristiti slični algoritmi uzorkovanja kao i za bootstrap koeficijente, za Y_+ je potrebno simulirati dodatnu varijabilnost - onu u odnosu na $x_+^T\beta$.

Predviđamo vrijednost $Y_+ = x_+^T\beta + \varepsilon_+$ koristeći vrijednost $\hat{Y}_+ = x_+^T\hat{\beta}$. Za slučajnu grešku ε_+ pretpostavljamo da je nezavisna te da dolazi iz iste distribucije kao i $\varepsilon_1, \ldots, \varepsilon_n$.

Za procjenu distribucije greške predikcije

$$\delta = \hat{Y}_+ - \hat{Y}_+ = (x_+^T\hat{\beta} - (x_+^T\beta + \varepsilon_+)) \tag{3.14}$$

koristimo distribuciju od

$$\delta^* = (x_+^T\hat{\beta}^* - (x_+^T\beta^* + \varepsilon_+^*)), \tag{3.15}$$

pri čemu je ε_+^* uzorkovan iz distribucije $\hat{G}$, a $\hat{\beta}^*$ je simuliran vektor procjena dobiven uzorkovanjem grešaka. Ovakav pristup pretpostavlja homoskedastičnost slučajnih grešaka. Prvi korak je izračunati modificirane reziduale, a nakon toga za svaki uzorkovani vektor dobivenih procijenjenih bootstrap koeficijenta $\hat{\beta}^*$ računamo 3.15. Algoritam je dan u nastavku.

Algorithm 4 Predikcija u linearnoj regresiji

1: **for** $r = 1, \ldots, R$ **do**
2: Simuliraj vrijednosti varijable odaziva y_r^* koristeći 2.22;
3: Izračunaj procjene koeficijenata $\beta_r^* = (X^TX)^{-1}X^Ty_r^*$;
4: **for** $m = 1, \ldots, M$ **do**
5: Nasumično izaberi $\varepsilon_{+,m}^*$ iz $r_1 - \bar{r}, \ldots, r_n - \bar{r}$
6: Izračunaj grešku predikcije $\delta_{r,m}^* = (x_+^T\hat{\beta}_r^* - (x_+^T\beta^* + \varepsilon_{+,m}^*))$.
7: **end for**
8: **end for**

Može se koristiti $M = 1$. Svrha uvođenja parametra M jest da RM bude dovoljno velik za procjenu svojstva δ^*.

Srednje kvadratna greška predikcije je procijenjena koristeći uzorkovanu srednje kvadratnu grešku

$$(RM)^{-1}\sum_{r,m}(\delta_{r,m}^* - \bar{\delta}^*)^2. \tag{3.16}$$

Ono što je poželjnije izračunati jest $(1-2\alpha)\%$ pouzdani interval za Y_+.

Za to su nam potrebni α i $(1-\alpha)$ stvarni kvantili predikcijske greške δ. Uz oznake spomenutih kvantila $a_\alpha, a_{1-\alpha}$, pouzdani intervali predikcije su sljedeći:

$$[\hat{y}_+ - a_{1-\alpha}, \quad \hat{y}_+ - a_\alpha].$$

No egzaktni kvantili, kao i funkcija distribucije, nepoznati su. Zato se služimo empirijskim kvantilima izračunatih procjena grešaka predikcije δ^* dobivenih iz 4, čiji redoslijed označavamo $\delta^*_{(1)} \leq \cdots \leq \delta^*_{(RM)}$. Dakle, bootstrap pouzdani intervali su dani kao

$$[\hat{y}_+ - \delta^*_{(RM+1)(1-\alpha)}, \quad \hat{y}_+ - \delta^*_{(RM+1)\alpha}].$$

Ovo jest analogan pristup kao bootstrap metodi uzorkovanja za pouzdane intervale.

Istaknimo na kraju još jednu sitnu modifikaciju koja omogućava nešto bolje rezultate. Ispostavlja se da je bolje u praksi koristiti studentiziranu predikcijsku grešku

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

pri čemu je S korijen iz srednje kvadratne greške reziduale linearne regresije. Objašnjenje je isto kao i kod bootstrap metoda za pouzdane intervale u poglavlju 5.2. i 5.4 literature [15]. Simulirane vrijednosti su

$$z^*_{r,m} = \frac{\delta^*_{r,m}}{s^*_r},$$

izračunate u drugom koraku algoritma 4. Traženi α i $(1-\alpha)$ kvantili od Z su procijenjeni koristeći ponovo empirijske kvantile, redom $z^*_{(RM+1)\alpha}$ i $z^*_{(RM+1)(1-\alpha)}$, pri čemu su $z^*_{(1)} \leq \cdots \leq z^*_{(RM)}$ poredani redom.

Konačno, studentizirani bootstrap predikcijski interval za Y_+ jest

$$[\hat{y}_+ - z^*_{(RM+1)(1-\alpha)}, \quad \hat{y}_+ - z^*_{(RM+1)\alpha}]. \tag{3.17}$$

Primjer 3.2.1. *Tablica 3.4 sadrži podatke o troškovima za reaktore s hlađenjem na običnu vodu*[2]*. Logaritam troška, mjeren u dolarima, jest varijabla odaziva. Sve ostale varijable su kovarijate, dok za kapacitet i N također uzimamo logaritam.*

Želimo procijeniti 95%*-tni interval pouzdanosti troška za iste vrijednosti kao i u posljednjem redu, uz promjenu datuma date* = 73*.*

[2]Podaci su dostupni i u sklopu R-ovog paketa *boot*.

Predikcija vrijednosti dobivena iz prilagođenog modela linearne regresije jest

$$x_+^T \hat{\beta} = 6.72,$$

uz srednje kvadratnu grešku reziduala

$$s = 0.159.$$

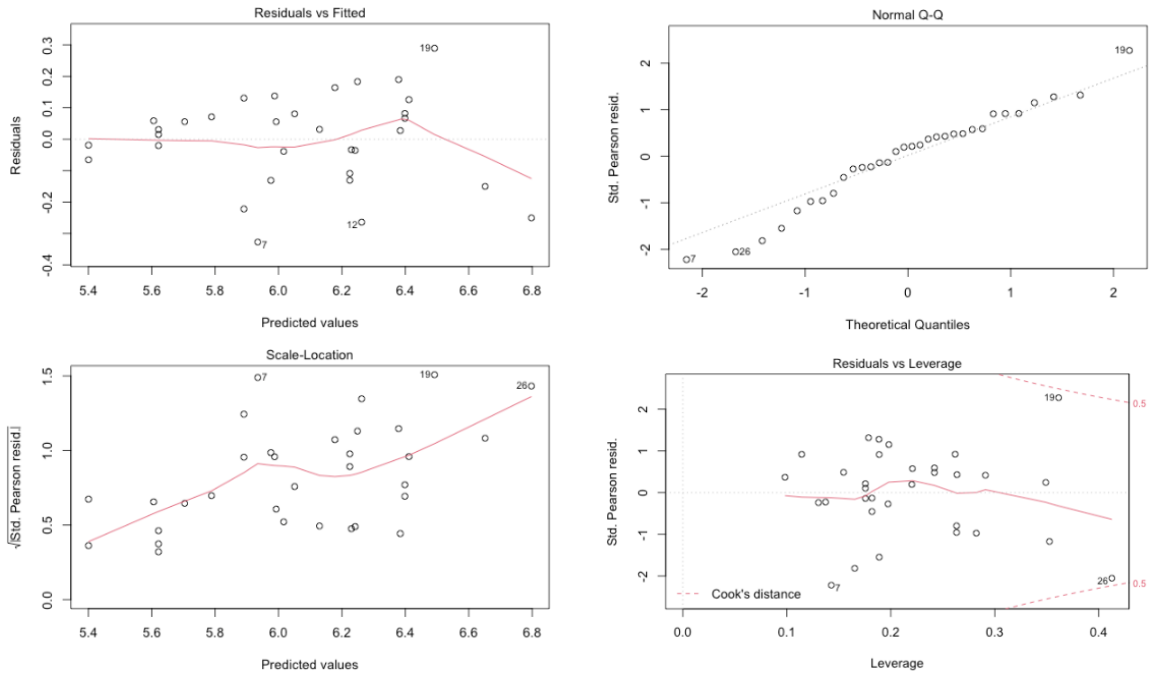
Uz $\alpha = 0.025$ *te korištenje koda 5, za* $R = 999$ *i* $M = 1$*, dobiveni su studentizirani kvantili*

$$z^*_{(25)} = -0.5743164$$
$$z^*_{(975)} = 0.5667300.$$

Dakle, prema formuli 3.17 predikcijski pouzdani interval troška iznosi

$$[469.9771, 1471.0480].$$

Kao dodatak na kraju ovog poglavlja ugrubo ćemo analizirati prilagodbu teorijskog višedimenzionalnog linearnog modela ovim podacima. Grafovi[3] *od interesa prikazani su na slici 3.3.*



Slika 3.3: Dijagnostika modela višedimenzionalne linearne regresije za podatke o reaktorima

[3]Prikaz je dobiven koristeći plot naredbu za *lm* model u R-u.

Usporedbom reziduala i prilagođenih vrijednosti vidimo nasumično raspršenje oko horizontalnog pravca $y = 0$*, što potvrđuje da je pretpostavka linearnosti bila ispravna. Jedino odstupanje se uočava kod opservacija s indeksom* $i = 7, 12, 19$*.*

Crtanjem q-q dijagrama uočena je zadovoljavajuća normalnost reziduala, iako su i ovdje uočena odstupanja opservacija $i = 7, 19$*, i dodatno za* $i = 26$*.*

Na scale-location grafu u donjem lijevom kutu, vidimo da pretpostavka homoskedastičnosti nije zadovoljena za ove podatke. Dobivena crvena linija nije ni blizu horizontalne, što potvrđuje nejednakost varijance grešaka, koju smo i očekivali. Sve ekstremne vrijednosti uočene prije, uočene su i na ovom prikazu.

Posljednji graf uspoređuje reziduale i poluge kako bismo utvrdili koje opservacije najviše utječu na naš model te isplati li ih se otkloniti iz modela. Vidimo da su to opservacije s indeksom 19 i 26. Micanjem tih opservacija iz podataka te ponovnom prilagodbom modelu vrijednost McFadden's R kvadrata[4] *raste sa* $R = 0.8568807$ *na* $R = 0.8808373$*. Dakle, možemo biti sigurni u odluku o otklanjanju tih vrijednosti.*

[4]McFadden's R kvadrat, poznatiji kao McFadden's R squared, mjera je koja koristi vjerodostojnost kao ocjenu za prilagodbu modelu. Formula glasi McFadden's R squared = 1 – (LogLikelihood(Prilagođeni model)/LogLikelihood(Model samo sa koeficijentom odsječka).

	cost	date	t1	t2	cap	pr	ne	ct	bw	cum.n	pt
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.50	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1090	0	0	0	0	1	0
20	490.88	68.92	16	59	1050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.80	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

Slika 3.4: Podaci o troškovima reaktora

Poglavlje 4

Praktični rezultati

U ovom poglavlju izloženi su rezultati dobiveni raznim simulacijama. Svaka simulacija provedena je koristeći sjeme *1234*[1]. Funkcije koje su korištene u programskom jeziku R priložene su kao dodatak na kraju rada. Simulacije su provedene za koeficijente jednostavne linearne regresije.

4.1 Osjetljivost bootstrapa na veličinu uzorka

Simuliramo uzorke iz uniformne razdiobe različitih duljina,

$$X \sim U(0, 100). \tag{4.1}$$

Broj iteracija držimo fiksnim,

$$R = 100. \tag{4.2}$$

Greške simuliramo iz normalne distribucije,

$$\varepsilon \sim N(0, 1). \tag{4.3}$$

Varijablu od interesa - Y simuliramo eksplicitno preko varijable X i slučajnih grešaka ε,

$$Y = 0.1 + 1.9X + \varepsilon. \tag{4.4}$$

Očito su vrijednosti koeficijenata jednake:

$$\beta_0 = 0.1 \tag{4.5}$$
$$\beta_1 = 1.9. \tag{4.6}$$

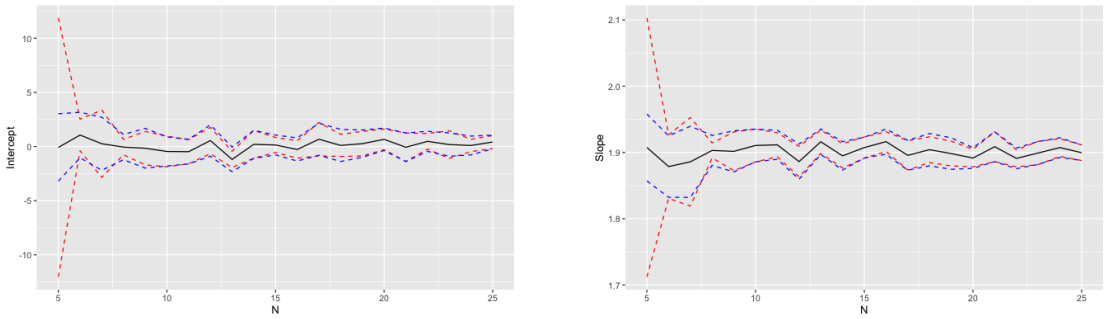
[1] Korištena naredba u R-u je set.seed(1234).

Mali broj podataka

Simulirani su uzorci duljina

$$n = 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25. \tag{4.7}$$

Dobiveni su sljedeći rezultati.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.1: 95%-tni pouzdani intervali bootstrap procjenitelja

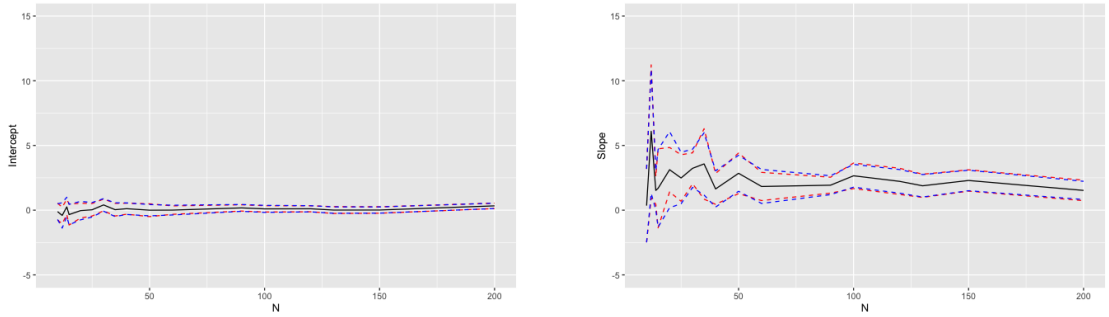
Plavom isprekidanom linijom prikazani su teorijski 95%-tni pouzdani intervali definirani u 1.28. Crvenom isprekidanom linijom prikazani su bootstrap pouzdani intervali dobiveni metodom uzorkovanja podataka. Za uzorke duljine manje od 8 bootstrap intervali su prilično široki. U svakom slučaju, stvarne vrijednosti $\beta_0 = 0.1$ i $\beta_1 = 1.9$ upadaju u dobivene intervale.

Veći uzorci

Ista stvar je provedena i za uzorke većih duljina,

$$n = 10, 12, 14, 15, 20, 25, 30, 35, 40, 50, 60, 90, 100, 120, 130, 150, 200. \tag{4.8}$$

Rezultati su prikazani na slici 4.2.
95%-tni pouzdani interval za koeficijent odsječka značajno je uži od pouzdanog intervala koeficijenta smjera. U oba slučaja se intervali sužuju kako $n \to \infty$. Također, zbog pripadnosti slučajnih grešaka normalnoj distribuciji bootstrap pouzdani intervali lijepo prate teorijske pouzdane intervale.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.2: 95%-tni pouzdani intervali bootstrap procjenitelja

4.2 Osjetljivost bootstrapa na broj iteracija

Sada simulacije provodimo za različit broj iteracija

$$R = 5, 10, 20, 30, 50, 70, 100, 200, 500, 1000, 2000, 3000, 10000.$$

Duljinu uzorka ne mijenjamo,

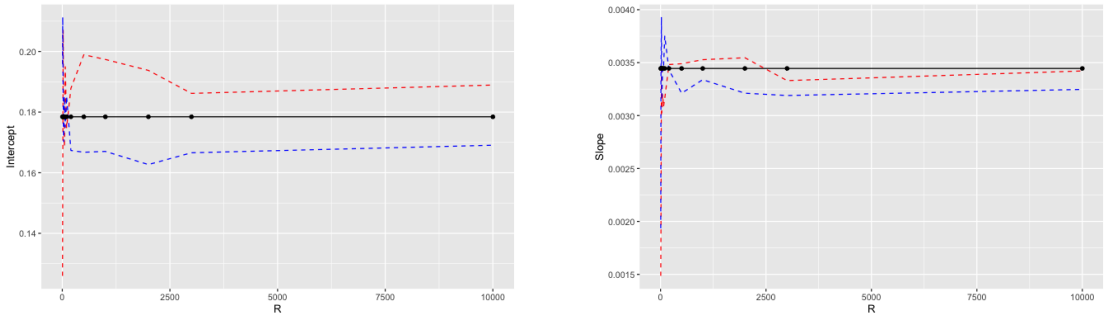
$$n = 100.$$

Ostale pretpostavke ostaju iste kao i u prethodnom poglavlju

$$\begin{aligned} X &\sim U(0, 100), \\ \varepsilon &\sim N(0, 1), \\ Y &= 0.1 + 1.9X + \varepsilon, \\ \beta_0 &= 0.1, \quad \beta_1 = 1.9. \end{aligned}$$

Rezultati su prikazani na slici4.3. Na x-osi označen je broj iteracija, dok je na y-osi prikazana standardna greška. Teorijska vrijednost označena je horizontalnom crnom linijom. Isprekidanom plavom linijom označena je vrijednost dobivena bootstrap metodom uzorkovanja grešaka, dok je crvenom isprekidanom linijom označena vrijednost dobivena uzorkovanjem podataka.
Za koeficijent odsječka vidimo da metoda uzorkovanja grešaka nudi najbolje rezultate, standardna greška je najmanja, čak manja i od teorijske. To nas ne čudi jer su greške homoskedastične, čime su pretpostavke za korištenje te metode - opravdane. Kod koeficijenta

smjera, za velik broj iteracija, obje bootstrap metode daju nešto bolje rezultate od teorijskih. Slična stvar kao i kod duljine uzorka, slučajna greška je manja, odnosno, intervali pouzdanosti su uži kako $r \to \infty$.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.3: Standardna greška bootstrap procjenitelja

4.3 Osjetljivost bootstrapa na distribuciju grešaka

Jedna od temeljnih pretpostavka za dokazivanje učinkovitosti bootstrap metoda jest pripadnost grešaka različitim distribucijama. Za normalne greške očekujemo poklapanje rezultata s teorijskim vrijednostima. To smo i uočili kod različitih duljina uzorka i različitog broja iteracija. U ovom poglavlju simuliraju se greške iz nekoliko poznatih statističkih distribucija. Parametri koji ostaju fiksni su:

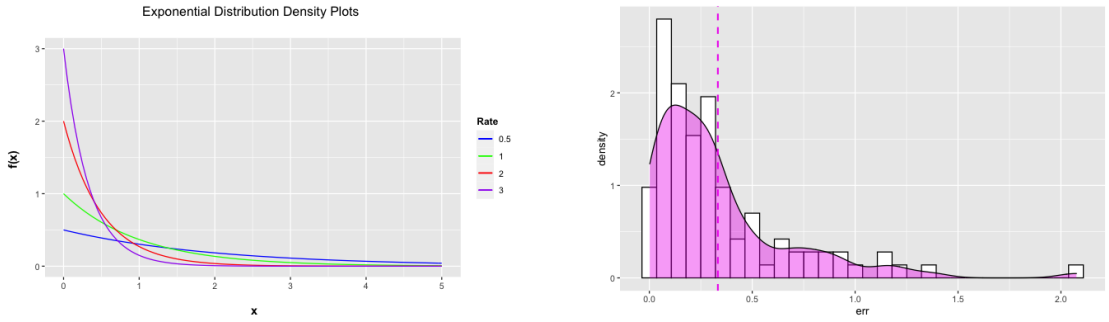
$$\begin{aligned} n &= 100, \\ X &\sim U(0, 100), \\ Y &= 0.1 + 1.9X + \varepsilon, \\ \beta_0 &= 0.1, \quad \beta_1 = 1.9. \end{aligned}$$

Eksponencijalna distribucija

Greške su simulirane iz eksponencijalne distribucije s parametrom $\lambda = 3$,

$$\varepsilon \sim Exp(3).$$

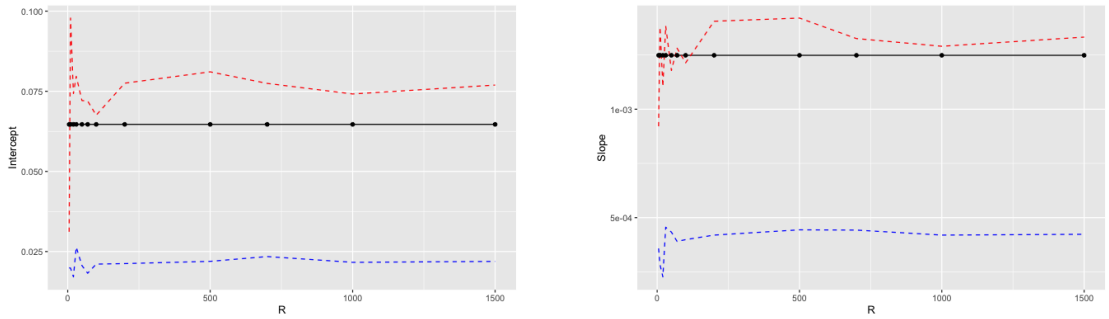
Na slici 4.4 lijevo prikazane su eksponencijalne funkcije gustoće s različitim parametrima λ. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.



(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.4: Eksponencijalna distribucija

Rezultati ovako simuliranih grešaka prikazani su na slici 4.5. Crnom horizontalnom linijom prikazana je teorijska standardna greška, plavom isprekidanom linijom standardna greška dobivena metodom uzorkovanja grešaka, a crvenom isprekidanom - standardna greška dobivena metodom uzorkovanja podataka.
Vidimo da najmanju standardnu grešku, za bilo koji broj iteracija R, daje metoda uzorkovanja grešaka. Metoda uzorkovanja podataka daje sličnu ocjenu, iako nešto goru od teorijskih vrijednosti.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

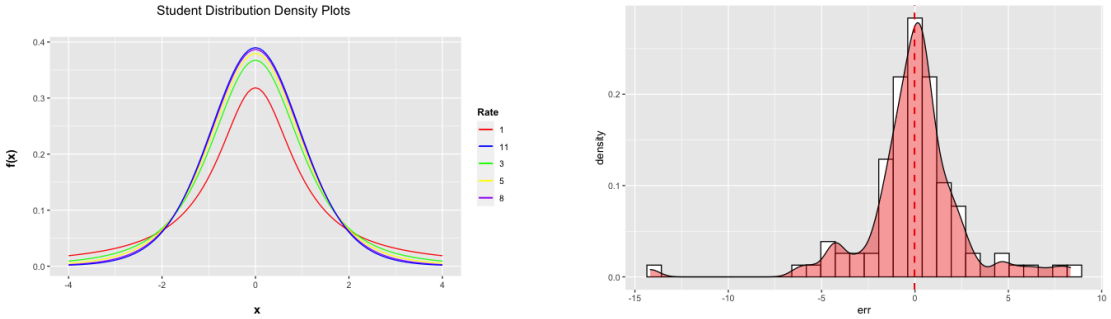
Slika 4.5: Standardna greška bootstrap procjenitelja

Studentova distribucija

Greške su simulirane iz studentove distribucije sa 1 stupnjem slobode,

$$\varepsilon \sim t_1. \tag{4.9}$$

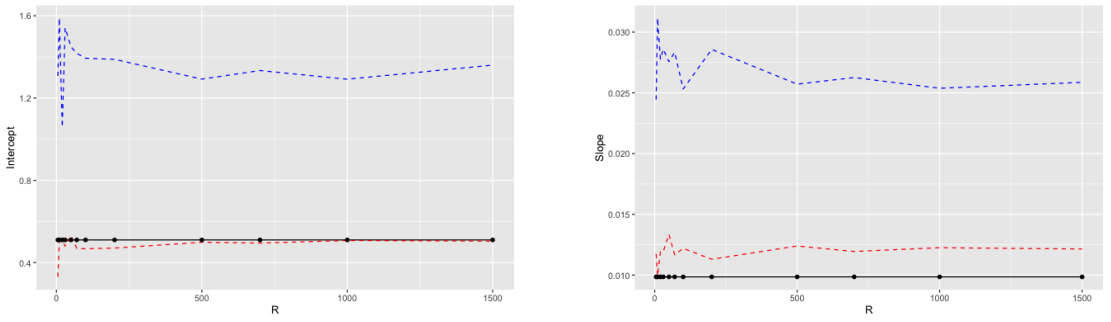
Na slici 4.6 lijevo prikazane su studentove funkcije gustoće s različitim stupnjevima slobode. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.



(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.6: Studentova distribucija

Rezultati ovako simuliranih grešaka prikazani su na slici 4.7. Crnom horizontalnom linijom prikazana je teorijska standardna greška, plavom isprekidanom linijom standardna greška dobivena metodom uzorkovanja grešaka, a crvenom isprekidanom - standardna greška dobivena metodom uzorkovanja podataka.
Vidimo da, iako su razlike male, bootstrap metode u ovom slučaju ne daju bolje rezultate od teorijskih. Metoda uzorkovanja grešaka u ovom slučaju ima najveću standardnu grešku za bilo koji broj iteracija R. Metoda uzorkovanja podataka daje sličan rezultat standardne greške kod koeficijenta odsječka, no za koeficijent smjera vidimo malo veću standardnu grešku od teorijske.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

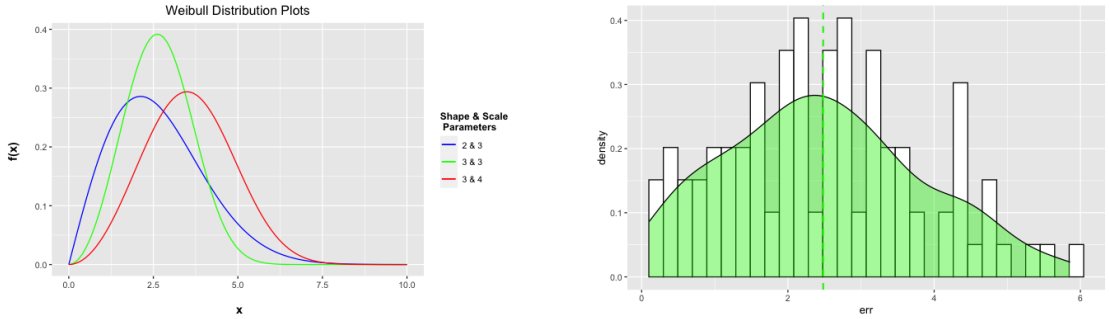
Slika 4.7: Standardna greška bootstrap procjenitelja

Weibullova distribucija

Greške su distribuirane iz Weibullove distribucije s parametrima $c = 3$ i $\alpha = 3$,

$$\varepsilon \sim W(3, 3). \tag{4.10}$$

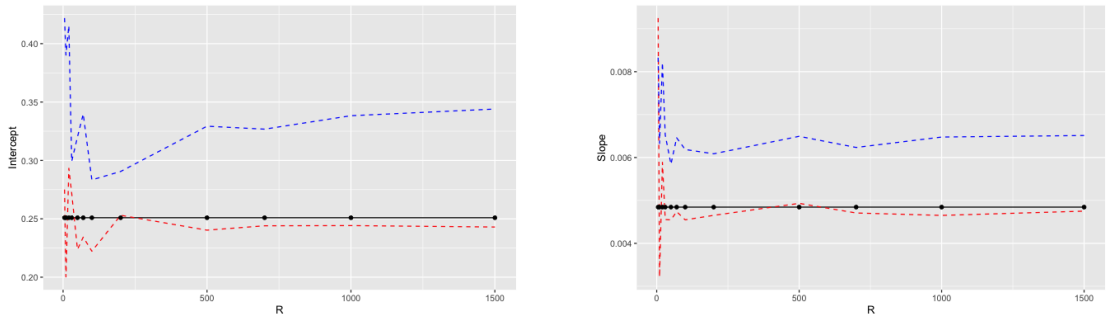
Na slici 4.8 lijevo prikazane su funkcije gustoće Weibullove distribucije s različitim parametrima. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.



(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.8: Weibullova distribucija

Dobiveni rezultati ovako simuliranih grešaka prikazani su na slici 4.9.
Slična stvar kao i kod studentove distribucije, bootstrap metoda ne daje značajno bolje rezultate od teorijskih. Razlike su male, pogotovo za koeficijent smjera. Metoda uzorkovanja podataka za velik broj iteracija daje slične rezultate kao što su teorijski.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

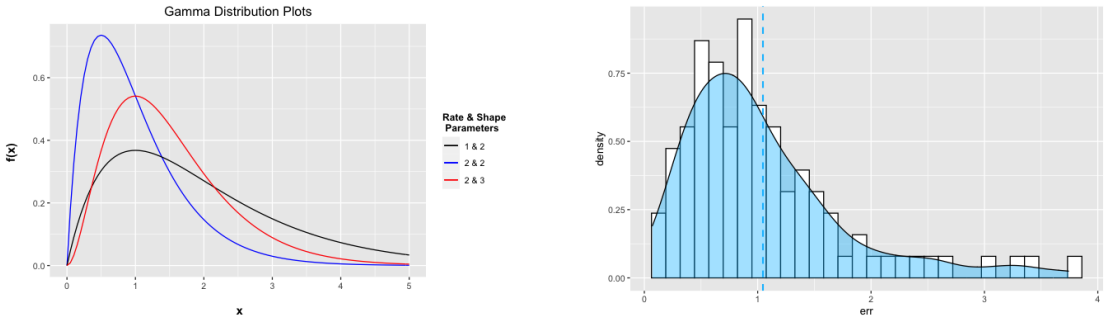
Slika 4.9: Standardna greška bootstrap procjenitelja

Gama distribucija

Greške su distribuirane iz gama distribucije s parametrima $\alpha = 2$ i $\beta = 2$,

$$\varepsilon \sim \Gamma(2, 2). \tag{4.11}$$

Na slici 4.8 lijevo prikazane su funkcije gustoće gama distribucije s različitim parametrima. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.

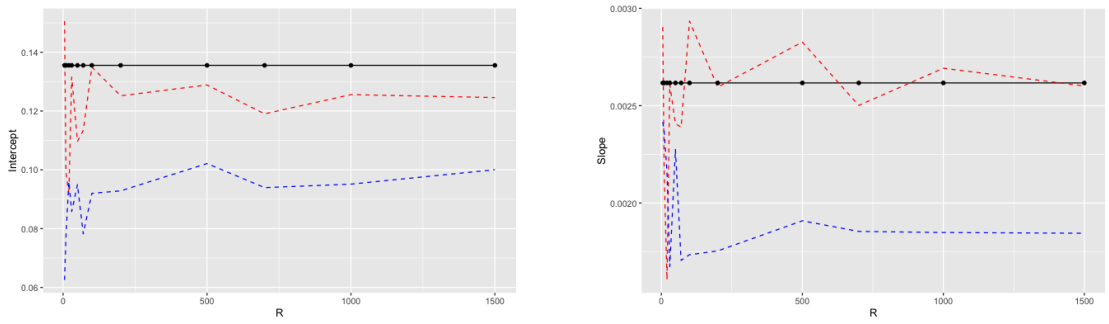


(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.10: Gama distribucija

Dobiveni rezultati ovako simuliranih grešaka prikazani su na slici 4.11.

Vidimo da za ovakvu distribuciju grešaka bootstrap metoda daje bolje rezultate. Metoda uzorkovanja grešaka daje značajno manju grešku i kod koeficijenta odsječka i kod koeficijenta smjera. Metoda uzorkovanja podataka je nešto bolja od teorijskih rezultata, ali gora od metoda uzorkovanja grešaka za koeficijent odsječka. Za koeficijent smjera daje zadovoljavajuće rezultate.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

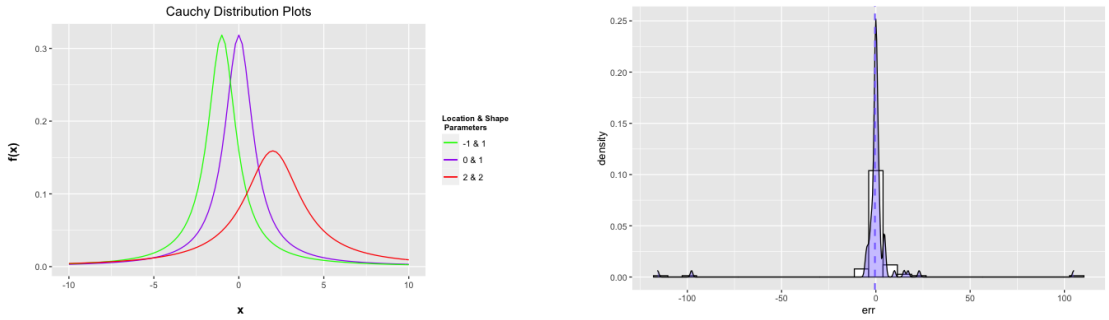
Slika 4.11: Standardna greška bootstrap procjenitelja

Cauchyjeva distribucija

Greške su simulirane iz Cauchyjeve distribucije s parametrima $x_0 = 0$ i $\gamma = 1$,

$$\varepsilon \sim C(1, 0). \tag{4.12}$$

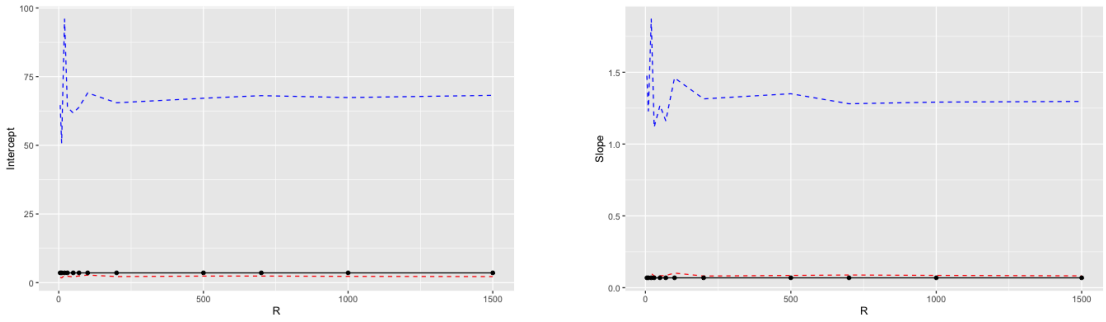
Na slici 4.12 lijevo prikazane su funkcije gustoće Cauchyjeve distribucije s različitim parametrima. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.



(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.12: Cauchyjeva distribucija

Dobiveni rezultati ovako simuliranih grešaka prikazani su na slici 4.13.
Metoda uzorkovanja grešaka u ovom slučaju ne funkcionira - dobivena standardna greška je velika, značajno viša od teorijske standardne greške. Metoda uzorkovanja podataka daje slične rezultate kao što su i teorijski.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

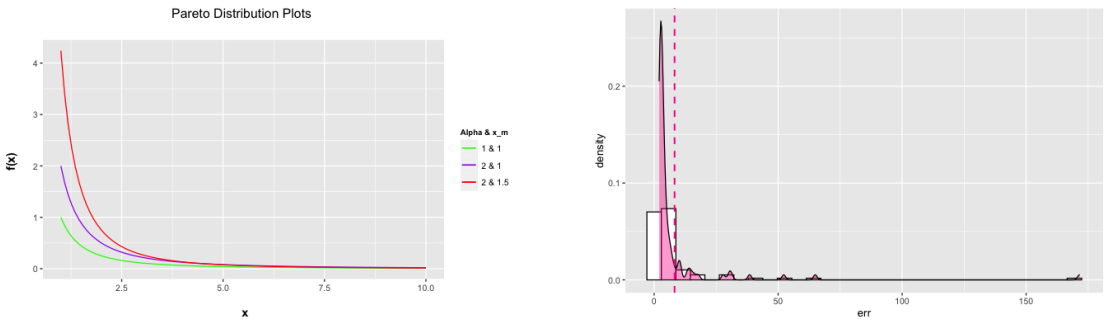
Slika 4.13: Standardna greška bootstrap procjenitelja

Paretova distribucija

Greške su distribuirane iz Paretove distribucije,

$$\varepsilon \sim Pareto(\frac{3}{2}, 2). \tag{4.13}$$

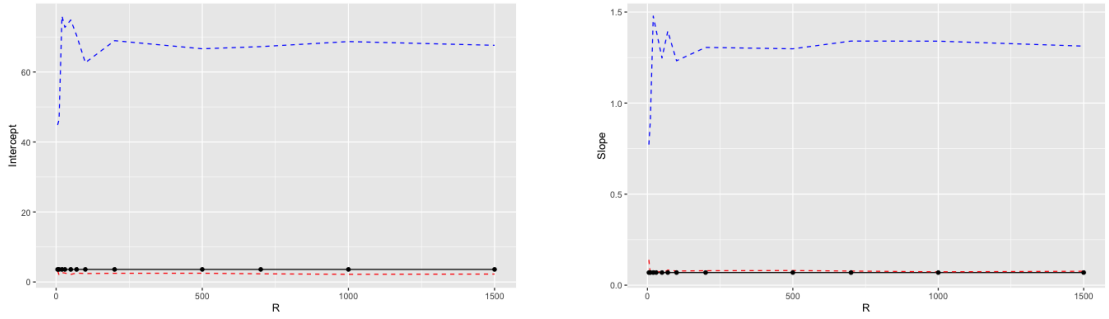
Na slici 4.14 lijevo prikazane su funkcije gustoće Paretove distribucije s različitim parametrima. S desne strane prikazan je histogram i procijenjena funkcija gustoće simuliranih grešaka.



(a) Funkcije gustoće **(b)** Histogram simuliranih grešaka

Slika 4.14: Pareto distribucija

Dobiveni rezultati ovako simuliranih grešaka prikazani su na slici 4.15.
Kao i kod Cauchyjeve distribucije, bootstrap metoda uzorkovanja grešaka ne daje dobre rezultate. Greška je itekako velika kod koeficijenta odsječka. Metoda uzorkovanja podataka u oba slučaja daje slične rezultate kao što su teorijski te nema značajnih razlika.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.15: Standardna greška bootstrap procjenitelja

Zaključak

U ovom poglavlju greške su simulirane iz šest različitih distribucija:

- eksponencijalne distribucije
- studentove distribucije
- Weibullove distribucije
- gama distribucije
- Cauchyjeve distribucije
- Paretove distribucije.

Najlošiji rezultati dobiveni su za posljednje dvije distribucije - Cauchyjevu distribuciju i Paretovu distribuciju. Očekivali smo da će upravo ovdje bootstrap dati najbolje rezultate s obzirom na to da obje distribucije pripadaju distribucijama teškog repa, tzv. *heavy tailed* distribucijama. Cauchyjeva distribucija ima dva "teška repa" - *two-tailed heavy*, dok Paretova ima jedan - *one-tailed heavy*. No u ovom slučaju razlog loših rezultata je u činjenici da varijanca ovih distribucija ne postoji, odnosno, $\mathrm{Var}(\varepsilon) = \infty$.

Za gama distribuciju, koja je *medium heavy*, te eksponencijalnu metoda uzorkovanja grešaka daje nešto bolje rezultate od metode uzorkovanja podataka, koja je prilično slična teorijskim podacima. Za distribucije koje djeluju više "normalno", odnosno pripadaju distribucijama lakog repa isto kao i normalna razdioba, kao što su studentova distribucija i weibullova distribucija, bootstrap metode ne nadmašuju teorijske rezultate - metoda uzorkovanja grešaka daje samo goru ocjenu standardne greške.[2]

[2]Klasifikacija funkcija distribucija s obzirom na "veličinu" repa preuzeta je iz [13]

4.4 Osjetljivost bootstrapa na varijancu grešaka

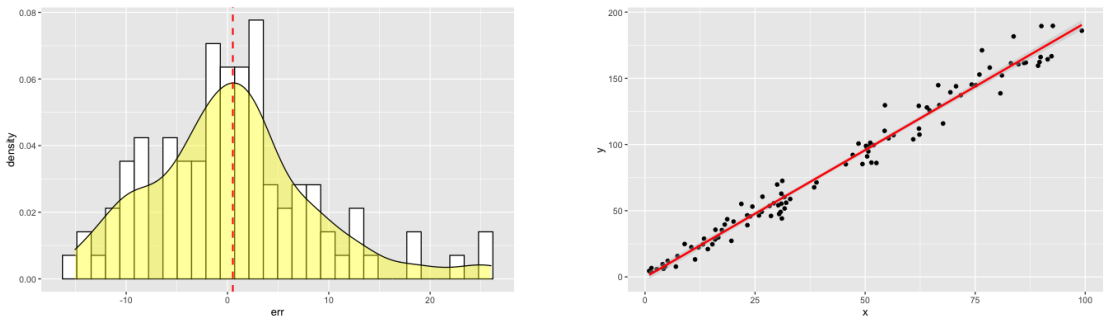
U gotovo svim simulacijama provedenim u prošlom poglavlju, varijanca grešaka je konstantna, definirana pripadnošću grešaka distribuciji određenoj parametrima. Sada ćemo se usredotočiti na simuliranje heteroskedastičnih, normalnih grešaka, čija se varijanca mijenja ovisno o vrijednostima slučajne varijable X. Razlikovat ćemo tri slučaja; blagu, srednju i jaku heteroskedastičnost.

Blaga heteroskedastičnost

Standardna devijacija slučajnih grešaka simulirana je korištenjem

$$sd_\varepsilon(x) = \sqrt{\frac{1}{2} + x}. \tag{4.14}$$

Na slici 4.16 prikazan je histogram simuliranih slučajnih grešaka te podaci s prilagođenim teorijskim modelom.



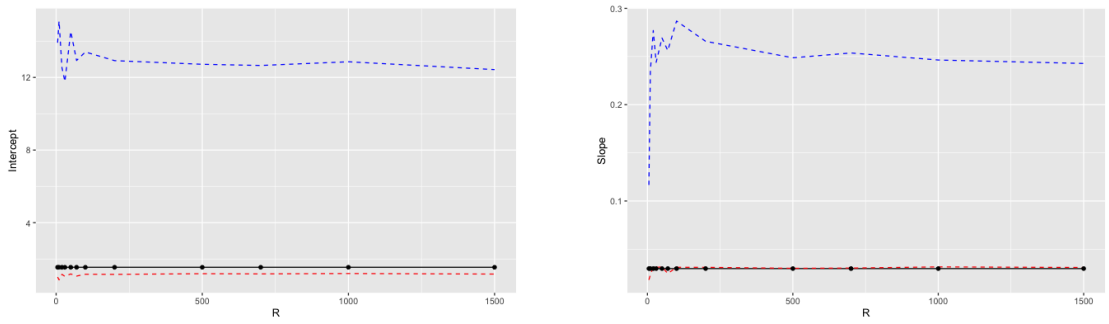
(a) Histogram simuliranih grešaka

(b) Prilagođen teorijski model

Slika 4.16: Blaga heteroskedastičnost normalnih grešaka

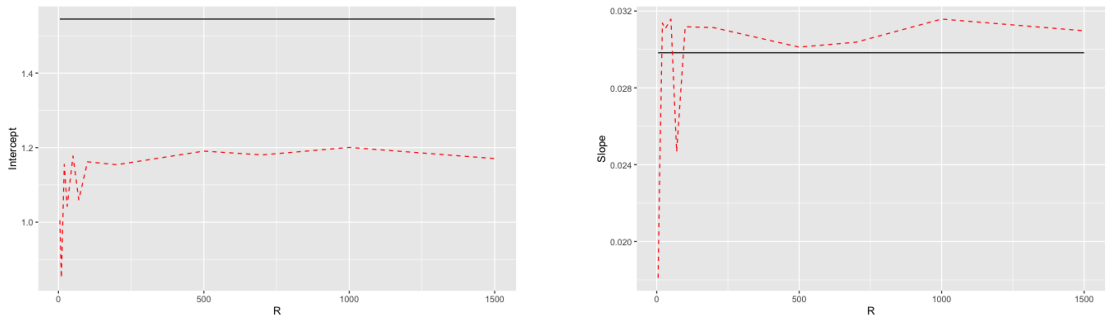
Rezultati su prikazani na slici 4.17 i 4.18. Na x-osi prikazan je broj iteracija, dok je na y-osi navedena standardna greška. Crnom linijom prikazana je teorijska vrijednost, plavom isprekidanom linijom vrijednost dobivena metodom uzorkovanja grešaka, a crvenom isprekidanom linijom vrijednost dobivena metodom uzorkovanja podataka. Prikaz u donjem redu uvećan je gornji prikaz kako bismo bolje uočili razliku između metode uzorkovanja podataka i teorijskih rezultata.

Metoda uzorkovanja grešaka vidljivo daje lošiji rezultat - standardna greška je veća za 12. Kod oba koeficijenta metoda uzorkovanja podataka lijepo prati teorijsku procjenu, dok za koeficijent odsječka čak i nadmašuje teoriju. Pobliža usporedba prikazana je na slici 4.18.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.17: Standardna greška bootstrap procjenitelja



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.18: Pobliži prikaz standardne greške za metodu uzorkovanja podataka

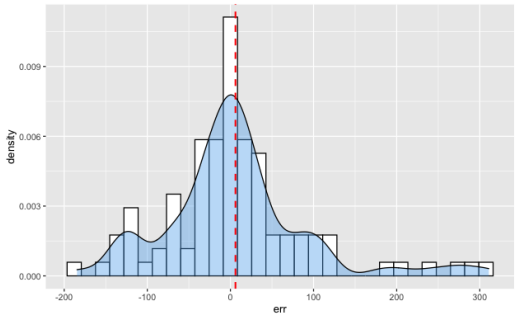
Srednja heteroskedastičnost

Standardna devijacija slučajnih grešaka simulirana je korištenjem

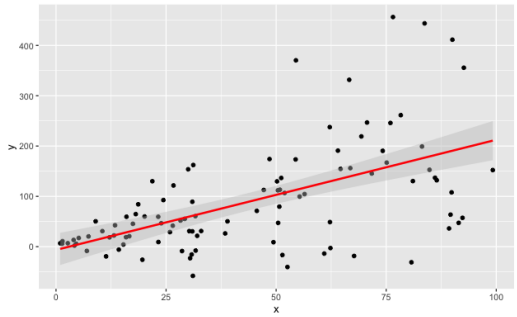
$$sd_{\varepsilon}(x) = 1 + x. \tag{4.15}$$

Na slici 4.19 prikazan je histogram simuliranih slučajnih grešaka te desno od njega podaci s prilagođenim teorijskim modelom. Podaci su prilično raspršeni, što već iz grafičkog prikaza daje naslutiti heteroskedastičnost.

Dobiveni rezultati su slični kao i prije. Bootstrap uzorkovanje grešaka i dalje ne daje zadovoljavajuće rezultate - standardna greška u ovom slučaju za koeficijent odsječka se kreće oko 1500. Bootstrap metoda uzorkovanja podataka daje bolju ocjenu standardne greške nego teorija u slučaju koeficijenta odsječka, dok je za koeficijent smjera lošija od teorijskih rezultata.

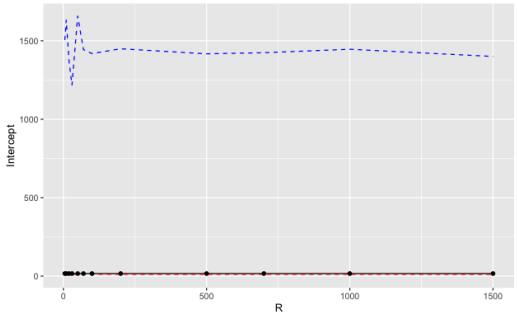


(a) Histogram simuliranih grešaka

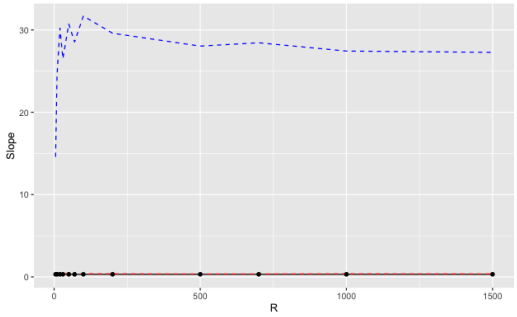


(b) Prilagođen teorijski model

Slika 4.19: Srednja heteroskedastičnost normalnih grešaka

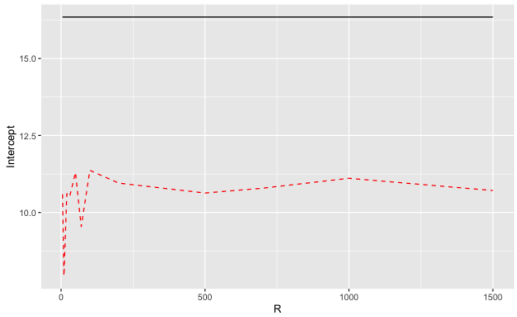


(a) Koeficijent odsječka

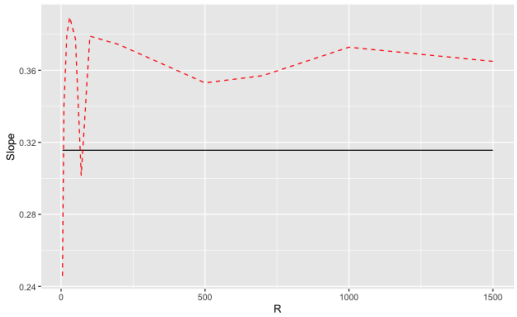


(b) Koeficijent smjera

Slika 4.20: Standardna greška bootstrap procjenitelja



(a) Koeficijent odsječka



(b) Koeficijent smjera

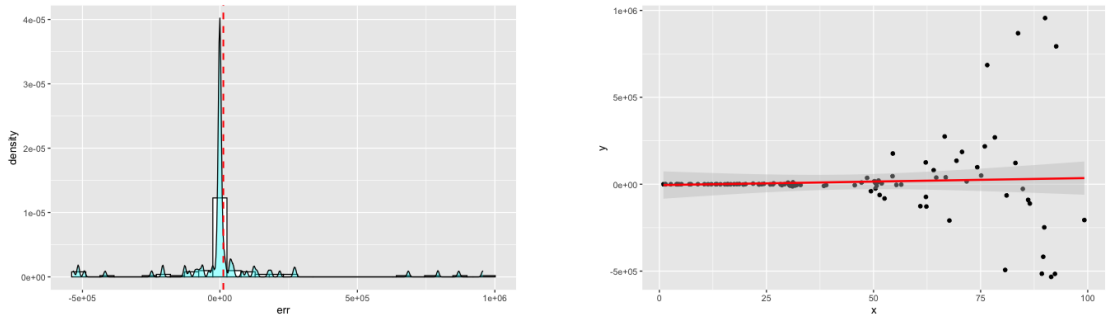
Slika 4.21: Pobliži prikaz standardne greške za metodu uzorkovanja podataka

Jaka heteroskedastičnost

Standardna devijacija slučajnih grešaka simulirana je korištenjem

$$sd_\varepsilon(x) = (1 + 1.08x)^5. \tag{4.16}$$

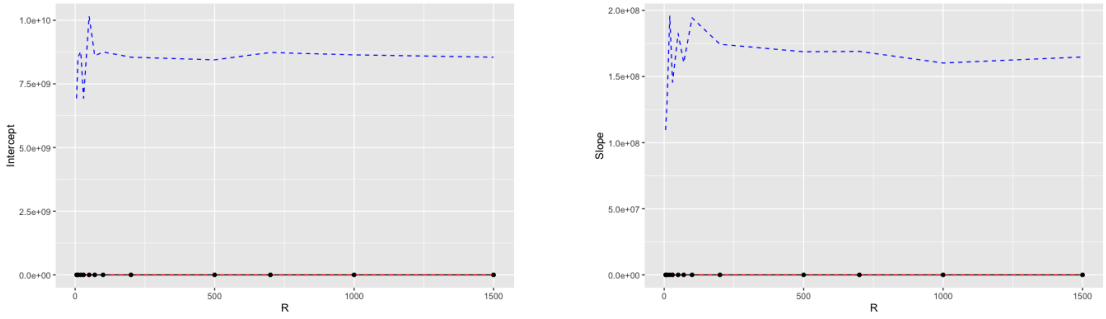
Na slici 4.22 prikazan je histogram simuliranih slučajnih grešaka te desno od njega podaci s prilagođenim teorijskim modelom. Podaci se raspršuju tek oko polovice. Opservacije slučajne varijable x koje su manje od 50 ne upućuju na nikakvu heteroskedastičnost, dok za podatke veće od 50 uočavamo velik skok u varijanci i raspršenje kako raste x. Ovo je očekivano s obzirom na to kako smo definirali funkciju standardne devijacije slučajnih grešaka.



(a) Histogram simuliranih grešaka **(b)** Prilagođen teorijski model

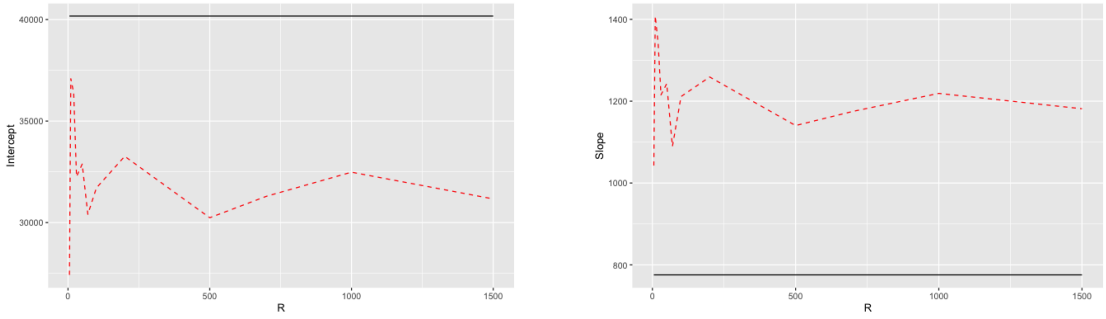
Slika 4.22: Jaka heteroskedastičnost normalnih grešaka

Rezultati su prikazani na slici 4.23. Metodi uzorkovanja grešaka u ovom slučaju nema spasa. Procjene standardne greške kreću se za red veličine 10^{10}. Metoda uzorkovanja podataka ovdje daje bolju ocjenu standardne greške za nekih 5000 u slučaju koeficijenta odsječka, dok kod koeficijenta smjera imamo preciznije teorijske rezultate.



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.23: Standardna greška bootstrap procjenitelja



(a) Koeficijent odsječka **(b)** Koeficijent smjera

Slika 4.24: Pobliži prikaz standardne greške za metodu uzorkovanja podataka

Zaključak

Korišteno je nekoliko različitih funkcija za simulaciju standardne devijacije slučajnih grešaka sa svrhom dobivanja heteroskedastičnosti. Uvjerili smo se da za takve distribucije grešaka bootstrap metoda uzorkovanja grešaka ne funkcionira. Štoviše, u nekim slučajevima procjene standardne greške su jako veliki brojevi. Budući da algoritam te metode pretpostavlja da greške dolaze iz jedne, jedinstveno određene distribucije, u slučaju ”puno” različitih distribucija grešaka ne daje dobre rezultate. Metoda uzorkovanje podataka ima više smisla u ovom slučaju jer nije toliko osjetljiva na varijancu grešaka i nudi veću robusnost u tom pogledu. Procjene su slične ili čak bolje od teorijskih.

Poglavlje 5

Dodatak - R-kod

Algoritam uzorkovanja grešaka za linearnu regresiju

```
Resampling_Errors <- function(R, x, y, modified_res, beta_0, beta_1){
  n <- length(x)
  est_slope = c()
  est_intercept = c()
  er_slope = c()
  er_intercept = c()
  for(i in 1:R){ #for each iteration
    x_new = c()
    y_new = c()
    for(j in 1:n){ #sample new data
      x_ = sample(x, 1)
      e_ = sample(modified_res - mean(modified_res), 1)
      y_ = beta_0 + beta_1 * x_ + e_
      x_new = append(x_new, x_)
      y_new = append(y_new, y_)
    }
    lsm_new = lsfit(x = x_new, y = y_new) #fit new least square model
    est_intercept = c(est_intercept, lsm_new$coefficients[1])
    est_slope = c(est_slope, lsm_new$coefficients[2])
    er_intercept = c(er_intercept, ls.diag(lsm_new)$std.err[1])
    er_slope = c(er_slope, ls.diag(lsm_new)$std.err[2])
  }
  return(data.frame(Intercept_Estimates = est_intercept,
                    Intercept_Errors = er_intercept,
                    Slope_Estimates = est_slope,
```

```
                    Slope_Errors = er_slope))
}
```

Algoritam uzorkovanja podataka za linearnu regresiju

```
Resampling_Cases <- function(R, x, y){
  n <- length(x)
  index = c(1:n)
  est_slope = c()
  est_intercept = c()
  er_slope = c()
  er_intercept = c()
  for(i in 1:R){ #for each iteration
    index_ = sample(index, n, replace = TRUE) #sample new data
    x_new = x[index_]
    y_new = y[index_]
    lsm_new = lsfit(x = x_new, y = y_new) #fit new model
    est_intercept = c(est_intercept, lsm_new$coefficients[1])
    est_slope = c(est_slope, lsm_new$coefficients[2])
    er_intercept = c(er_intercept, ls.diag(lsm_new)$std.err[1])
    er_slope = c(er_slope, ls.diag(lsm_new)$std.err[2])
  }
  return(data.frame(Intercept_Estimates = est_intercept,
                    Intercept_Errors = er_intercept,
                    Slope_Estimates = est_slope,
                    Slope_Errors = er_slope))
}
```

Uzorkovanje grešaka za multidimenzionalnu linearnu regresiju

```
    Resampling_Errors <- function(R, X, y, modified_res, beta){
  n = nrow(X)
  M =  matrix(, nrow = 0, ncol = ncol(X))
  for(i in 1:R){
    row_index = sample(1:n, n, replace = TRUE)
    e_new = sample(modified_res - mean(modified_res), n)
    y_new = sum(beta * X) + e_new
    lsm_new = lsfit(X, y = y_new, intercept = FALSE)
    M = rbind(M, t(ls.diag(lsm_new)$std.err))
```

```
  }
  return(colMeans(M))
}
```

Uzorkovanje podataka za multidimenzionalnu linearnu regresiju

```
    Resampling_Cases <- function(R, X, y){
  n = nrow(X)
  M =  matrix(, nrow = 0, ncol = ncol(X))

  for(i in 1:R){
    row_index = sample(1:n, n, replace = TRUE)
    X_new = X[row_index, ]
    y_new = y[row_index]
    lsm_new = lsfit(x = X_new, y = y_new, intercept = FALSE)
    M = rbind(M, t(ls.diag(lsm_new)$std.err))
  }
  return(colMeans(M))
}
```

Uzorkovanje podataka - pamćenje svojstvene vrijednosti

```
Resampling_Cases_Eigen <- function(R, X, y){
  n = nrow(X)
  M =  matrix(, nrow = 0, ncol = 6)
  ones = rep(1, n)
  for(i in 1:R){
    row_index = sample(1:n, n, replace = TRUE)
    X_new = X[row_index, ]
    y_new = y[row_index]
    X_new_ = cbind(ones, X_new)
    lambda = min(eigen(t(X_new_)%*% X_new_)$values)
    model = glm(y_new ~ X_new)
    errors = summary(model)$coefficients[, 2]
    M = rbind(M, c(errors, lambda))
  }
  return(data.frame(M))
}
```

Uzorkovanje podataka korištenjem *boota*

Kodovi izloženi na ovoj stranici napisani su za primjer podatke o betonu.

```
fit_model <- function(data){
    fit = glm(y ~ x1+x2+x3+x4, data = data)
    c(coef(fit))
}
boot_fun = function(data, i) fit_model(data[i, ])
boot(cement, boot_fun, R = 1000)
```

Uzorkovanje grešaka korištenjem *boota*

```
fit_model <- function(data){
fit = glm(y ~ x1+x2+x3+x4, data = data)
c(coef(fit))
}
boot_fun_erors = function(data, i){
    d = data
    d$y = d$fit + d$res[i]
    fit_model(d)
}
boot(tmp, boot_fun_erors, R = 1000)
```

Predikcija

```
nuclear_model = glm(log(cost) ~
pt+ct+ne+date+log(cap)+log(cum.n), data = nuclear)
nuclear_diag = glm.diag(nuclear_model)

tmp = data.frame(nuclear, fit = fitted(nuclear_model),
res = nuclear_diag$res * nuclear_diag$sd)

x_predict = tmp[32, c(11, 8,7, 5, 10)]
x_predict$date = 73
x_predict$fit = predict(nuclear_model, x_predict)

boot_pred_function = function(data, i, i.p, d.p){
d = data
d$cost = exp(d$fit + d$res[i]) #calculate y, i is the sampling
```

```
  d.glm = glm(log(cost) ~ pt+ct+ne+date+log(cap)+log(cum.n),
  data = d)
  predict(d.glm, d.p) - (d.p$fit+d$res[i.p])
  }

  nuclear_boot = boot(tmp, boot_pred_function, R = 599,
  m=1, d.p=x_predict)

  as.vector(exp(x_predict$fit -
  quantile(nuclear_boot$t, c(0.975, 0.025)))
```

Osjetljivost bootstrapa na veličinu uzorka

```
  sample_size_comparasion <- function(N, R){
upper_int = c()
lower_int = c()
upper_slope = c()
lower_slope = c()
teory_int = c()
teory_slope = c()
teory_upper_int = c()
teory_upper_slope = c()
teory_lower_int = c()
teory_lower_slope = c()

for(n in N){
  x = runif(n, 0, 100)
  err = rnorm(n, 0, 1)
  y = 0.1 + 1.9 * x + err
  df = data.frame(x, y)
  main_model = lm(y ~ x, data = df)
  boot_beta <- replicate(R, {
    index = sample(n, n, replace = TRUE)
    fit_boot = lm(y ~ x, data = df[index, ])
    coef(fit_boot)
    })
  se_int = sd(boot_beta[1,])
  se_slope = sd(boot_beta[2,])
```

```
    upper_int = c(upper_int, coef(main_model)[1] +
    qnorm(1-.05/2) * se_int)
    upper_slope = c(upper_slope, coef(main_model)[2] +
    qnorm(1-.05/2) * se_slope)

    lower_int = c(lower_int, coef(main_model)[1] -
    qnorm(1-.05/2) * se_int)
    lower_slope = c(lower_slope, coef(main_model)[2] -
    qnorm(1-.05/2) * se_slope)
    teory_int = c(teory_int, coef(main_model)[1])
    teory_slope = c(teory_slope, coef(main_model)[2])

    teory_upper_int = c(teory_upper_int, confint(main_model)[1,2])
    teory_upper_slope = c(teory_upper_slope, confint(main_model)[2,2])
    teory_lower_int = c(teory_lower_int, confint(main_model)[1,1])
    teory_lower_slope = c(teory_lower_slope, confint(main_model)[2,1])


  }

  return(results = data.frame(Lower_int = lower_int,
  Upper_int = upper_int,
  Lower_slope = lower_slope,
  Upper_slope = upper_slope,
  Teory_int = teory_int,
  Teory_slope = teory_slope,
  Teory_lower_int =  teory_lower_int,
  Teory_upper_int = teory_upper_int,
  Teory_lower_slope = teory_lower_slope,
  Teory_upper_slope = teory_upper_slope,
  row.names = N))
}
```

Osjetljivost bootstrapa na broj iteracija

```
    iteration_number_comparasion <- function(R){

  teory_err_int = c()
  res_cas_err_int  = c()
```

```
res_err_err_int  = c()
teory_err_slope = c()
res_cas_err_slope  = c()
res_err_err_slope  = c()

n = 100
x = runif(n, 0, 100)
err = rnorm(n, 0, 1)
y = 0.1 + 1.9 * x + err
df = data.frame(x, y)
main_model = lm(y ~ x, data = df)
residuals = resid(main_model)
residuals = (residuals - mean(residuals)) * sd(residuals)

for(r in R){
  boot_beta_res_cases <- replicate(r, {
    index = sample(n, n, replace = TRUE)
    fit_boot = lm(y ~ x, data = df[index, ])
    coef(fit_boot)
    })

  boot_beta_res_err <- replicate(r, {
    index = sample(n, n, replace = TRUE)
    y_ = fitted(main_model) + residuals[index]
    fit_boot = lm(y_ ~ x)
    coef(fit_boot)
    })

  teory_err_int = c(teory_err_int,
  summary(main_model)$coefficients[,2][1])
  res_cas_err_int  = c(res_cas_err_int,
  sd(boot_beta_res_cases [1,]))
  res_err_err_int  = c(res_err_err_int,
  sd(boot_beta_res_err [1,]))
  teory_err_slope = c(teory_err_slope,
  summary(main_model)$coefficients[,2][2])
  res_cas_err_slope  = c(res_cas_err_slope,
  sd(boot_beta_res_cases [2,]))
  res_err_err_slope  = c(res_err_err_slope,
```

```
    sd(boot_beta_res_err [2,]))


  }

  return(results = data.frame(teory_err_int,
  res_cas_err_int,
  res_err_err_int,
  teory_err_slope,
  res_cas_err_slope,
  res_err_err_slope,
  row.names = R))
}
```

Bibliografija

[1] *Data.gov.kr*, studeni 2022., https://www.data.go.kr/data/15007122/fileData.do.

[2] *Bootstrap and Linear regression*, https://www.maxturgeon.ca/f21-stat3150/slides/bootstrap-linreg.pdf, preuzeto 28.11.2022.

[3] *Formula za varijancu u linearnoj regresiji*, https://stats.stackexchange.com/questions/115011/in-simple-linear-regression-where-does-the-\formula-for-the-variance-of-the-resi, preuzeto 29.1.2023.

[4] Vedat Akgiray, *Conditional Heteroscedasticity in Time Series of Stock Returns: Evidence and Forecasts*, The Journal of Business (1989).

[5] Robert B. Ash, *Real Analysis and Probability*, Academic Press, 1972.

[6] Daniel K. Baissa1 i Carlisle Rainey, *When BLUE is not best: non-normal errors and the linear model*, Political Science Research and Methods (2020).

[7] F. Cribari-Neto i S. G. Zarkos, *Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis*, Econometrics Reviews (2007).

[8] Zlatko Drmač, *Numerička analiza, Predavanja i vježbe*, Zagreb, 2003.

[9] B. Efron i R. Tibshiran, *An Introduction to the Bootstrap*, Chapman and Hall, 1993.

[10] Anja F. Ernst i Casper J. Albers, *Regression assumptions in clinical psychology research practice—a systematic review of common misconceptions*, PubMed Central (2017).

[11] Emmanuel Flachaire, *Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap*, Computational Statistics and Data Analysis (2005).

[12] S. Foss i D. Korshunov, *An Introduction to Heavy-Tailed and Subexponential Distributions*, New York, 2013.

[13] Leigh J. Halliwell, *Classifying the Tails of Loss Distributions*, Casualty Actuarial Society E-Forum (2013).

[14] Qiyang Han i Jon A. Wellner, *Convergence rates of least squares regression estimators with heavy-tailed errors*, The Annals of Statistics (2019).

[15] Hinkley i Davison, *Bootstrap Methods and their Application*, Cambridge University Press, 1997.

[16] Miljenko Huzak, *Matematička statistika, Predavanja*, Zagreb, 2020./2021.

[17] M. Kutner i C.Nachtsheim, *Applied Linear Statistical Models, 5th Edition*, New York, 2004.

[18] Petra Lazić, *Linearni modeli 2*, `https://web.math.pmf.unizg.hr/nastava/statpr2/materijali/sp2_vjezbe3.pdf`, preuzeto 15.12.2022.

[19] Edmore Ranganai, *On studentized residuals in the quantile regression framework*, Springerplus (2016).

[20] Nikola Sarapa, *Teorija vjerojatnosti*, Školska knjiga, Zagreb, 2002.

[21] Henk Talma i Paula van Dommelen, *The world's tallest nation has stopped growing taller: the height of Dutch children from 1955 to 2009*, Pediatr Res (2022), `https://pubmed.ncbi.nlm.nih.gov/23222908/`.

[22] Z Xu, *Waist-to-height ratio is the best indicator for undiagnosed type 2 diabetes*, Diabetic Medicine (2013), `https://pubmed.ncbi.nlm.nih.gov/23444984/`.

Sažetak

Linearna regresija najčešći je oblik regresijske analize. Ako su pretpostavke normalnosti, homogenosti i nezavisnosti slučajnih grešaka ispunjene, metoda najmanjih kvadrata daje egzaktne rezultate. U praksi su pretpostavke često oslabljene ili čak nedostižne, čime se traži alternativa u procjeni koeficijenata. Metode ponovnog uzorkovanja koje se zasnivaju na zakonu velikih brojeva nameću se kao moguće potencijalno rješenje.

U ovom radu proučava se *bootstrap* metoda, koja polazi od ideje da statistički uzorak već sadrži sve dostupne informacije o nekoj funkciji distribucije F te ponovnim uzorkovanjem dobiva se uzorak koji pripada toj istoj distribuciji. U središtu interesa su dva algoritma - metoda uzorkovanja podataka i metoda uzorkovanja grešaka.

U prvom poglavlju iznose se osnovni pojmovi teorije vjerojatnosti te se predstavlja *bootstrap* metoda. U drugom poglavlju definiran je jednostavan model linearne regresije te je iznesena ideja *bootstrap* metode za distribuciju procjenitelja. Treće poglavlje je proširivanje drugog poglavlja na višedimenzionalan slučaj. Također, predstavljen je pojam predikcije te uloga *bootstrap* metode u tom pogledu. U četvrtom poglavlju izloženi su praktični rezultati dobiveni raznim simulacijama u programskom jeziku R. Dodatak na kraju rada je peto poglavlje, gdje se nalazi kod korišten za provođenje simulacija.

Summary

Linear regression is the most common form of regression analysis. If the assumptions of normality, homogeneity and independence of random errors are satisfied, the least square method gives exact results. In practice, assumptions are often difficult to reach or even elusive, which requires looking for an alternative. Resampling methods, based on the law of large numbers, are offered as a possible potential solution.

In this paper, the bootstrap method is studied, which is based on the idea that a statistical sample already contains all available information about a distribution function F, so by resampling, we get a sample that belongs to that same distribution. Two algorithms are particularly discussed – the data resampling method and the errors resampling method.

In the first chapter, the basic concepts of probability theory and the bootstrap method are presented. In the second chapter, a simple linear regression model is defined and the idea of the bootstrap method for the distribution of estimators is exposed. The third chapter is an extension of the second chapter to a multidimensional case. Also, the concept of prediction and the role of the bootstrap method in this respect are presented. In the fourth chapter, the practical results obtained by conducting various simulations in the R programming language are exposed. The code used to perform the simulations is presented at the end of the paper, in the fifth chapter.

Životopis

Rođena sam 22. rujna 1998. godine u Varaždinu. Treću osnovnu školu Varaždin završavam 2013. godine. Iste godine upisujem Prvu gimnaziju Varaždin, a dvije godine nakon toga Centar izvrsnosti Varaždinske županije iz matematike. Tijekom srednjoškolskog obrazovanja sudjelujem na matematičkim natjecanjima te 2017. stječem titulu županijske prvakinje u matematici. Iste godine upisujem Prirodoslovno-matematički fakultet Sveučilišta u Zagrebu, gdje nastavljam svoje obrazovanje. Trenutačno sam zaposlena u *Medtronic Adriaticu* na poziciji specijalista za analize.