Domača naloga 3

Neža Kržan, Tom Rupnik Medjedovič

1 Cilji naloge

Želiva preučiti uporabo metode ponovnega vzorčenja *bootstrap*, v primeru ko klasičnimi testom ne moremo popolno zaupati zaradi kršenja predpostavk. Analizirati želiva primer, kjer klasična statistična metoda, v najinem primeru je to linearna regresija, ni popolnoma primerna za ocenjevanje koeficientov napovednih spremenljivk in njihovih intervalov zaupanja, zaradi kršitve predpostavk statističnega modela(tj. linearna regresija).

Generirala bova podatke, na katerih bova izračunala intervale zaupanja koeficientov linearne regresije.

Za uporabno linearne regresije je potrebno izpolniti določene predpostavke, da imamo veljaven model. Te so:

- linearna odvisnost: *obstoj linearne povezanosti med napovednimi (pojasnjevalnimi) spremenljivkami in odzivno (ciljno) spremenljivko*,
- normalna porazdeljenost napak in neodvisnost napak,
- homoskedastičnost: *varianca napak mora biti konstantna*,
- brez popolne multikolinearnosti: *neodvisne spremenljivke ne smejo biti preveč povezane med seboj*,
- zadostno število podatkov.

Podatke bova generirala tako, da bodo nekatere izmed teh predpostavk kršene(opisano v naslednjem poglavju) ter bova med seboj primerjala intervale zaupanja dobljene z linearno regresijo (`lm`) in metodo ponovnega vzorčenja(*bootstrap*).

2 Generiranje podatkov

Podatke sva generirala tako, da je v linearnem regresiji kršena predpostavka homoskedastičnost (konstantna varianca napak). Enačba, ki sva jo uporabila za generiranje odzivne spremenljivke je enaka:

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \epsilon_i$$

pri čemer:

- β_0-konstanta enaka 100,
- β_1-koeficient spremenljivke x_1 enak 3,
- β_2-koeficient spremenljivke x_2 enak 2,
- ϵ-napaka, ki generirana iz porazdelitve $N(0, x_1 \cdot \alpha)$ (α določa povezanost s spremenljivko x_1),

torej

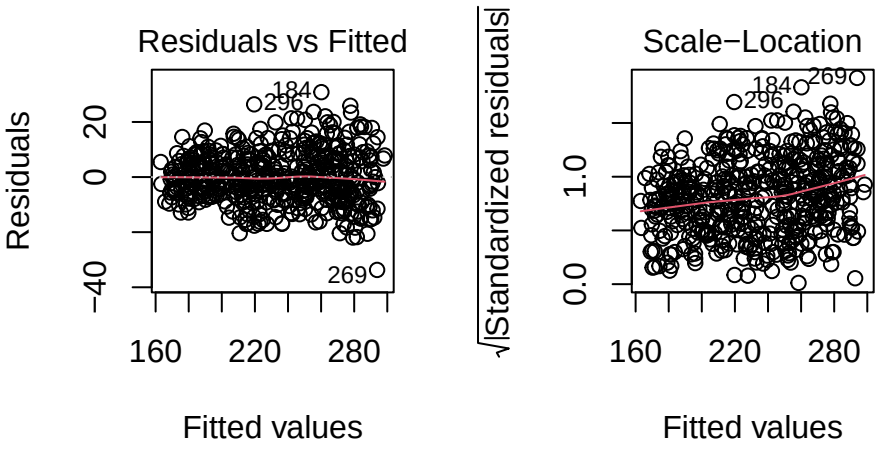
$$y_i = 100 + 3 \cdot x_{1i} + 2 \cdot x_{2i} + \epsilon_i.$$

Kot sva že omenila bova spreminjala faktor α in s tem kršila predpostavko konstantne variance napak, saj se ta z večanjem vrednosti x_1 povečuje ali pa je konstantna. Ta zavzame vrednosti $\alpha \in \{0.6, 1, 1.2\}$.

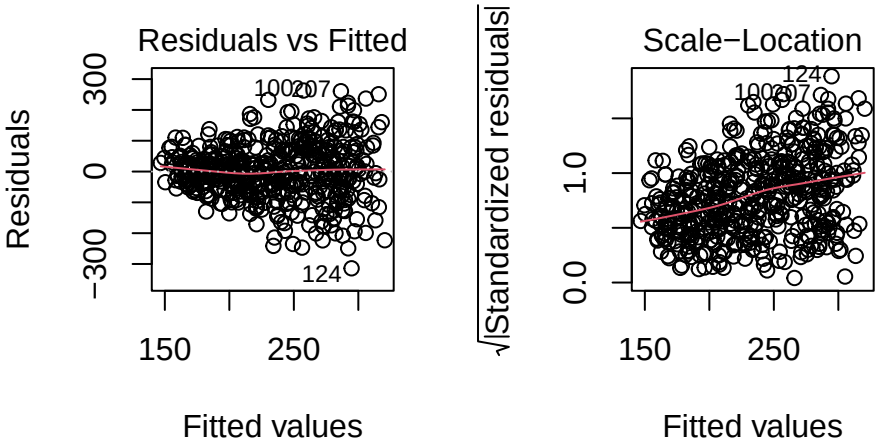
Velikost vzorca pa enak $n = (20, 200, 500)$, ker naju zanima kako se bo bootstrap metoda obnesla tudi na zelo majhnem vzorcu - s tem je tudi kršena predpostavka o dovolj velikem vzorcu pri linearni regresiji.

Pri generiranju posameznih vrednosti v enačbi linearne regresije(x_{1i}, x_{2i}) sva se odločila za generiranje iz enakomerne porazdelitve, in sicer $x_{1i} \sim Unif(20, 60)$ ter $x_{2i} \sim Unif(2, 10)$, torej, da imamo v podatkih majhne vrednosti in nekoliko večje.

Narišimo grafe ostankov za nekaj kombinacij faktorjev, da se prepričamo o kršenju predpostavk, velikost vzorca je v *vseh* primerih nastavljena na 500.



Slika 1: Grafi ostankov pri paramatrih alpha=0,6.



Slika 2: Grafi ostankov pri paramatrih alpha=1,2.

V obeh primerih lahko na desnem grafu opazimo, da se z večanjem vrednosti povečuje tudi variabilnost napak (naraščajoč trend). To lahko opazimo, tudi iz levega grafa, saj se od leve proti desni s povečevanjem vrednosti, povečuje tudi variabilnost ostankov. Prav tako je vidna razlika, ko povečamo vrednost parametra α, saj se vrednosti ostankov povečajo (variabilnost se poveča).

3 Klasični test in metoda ponovnega vzorčenja

Pri ponovnem vzorčenju bova uporabila metodo bootstrap, pri kateri bo število bootstrap vzorcev enako $m = 1000$.

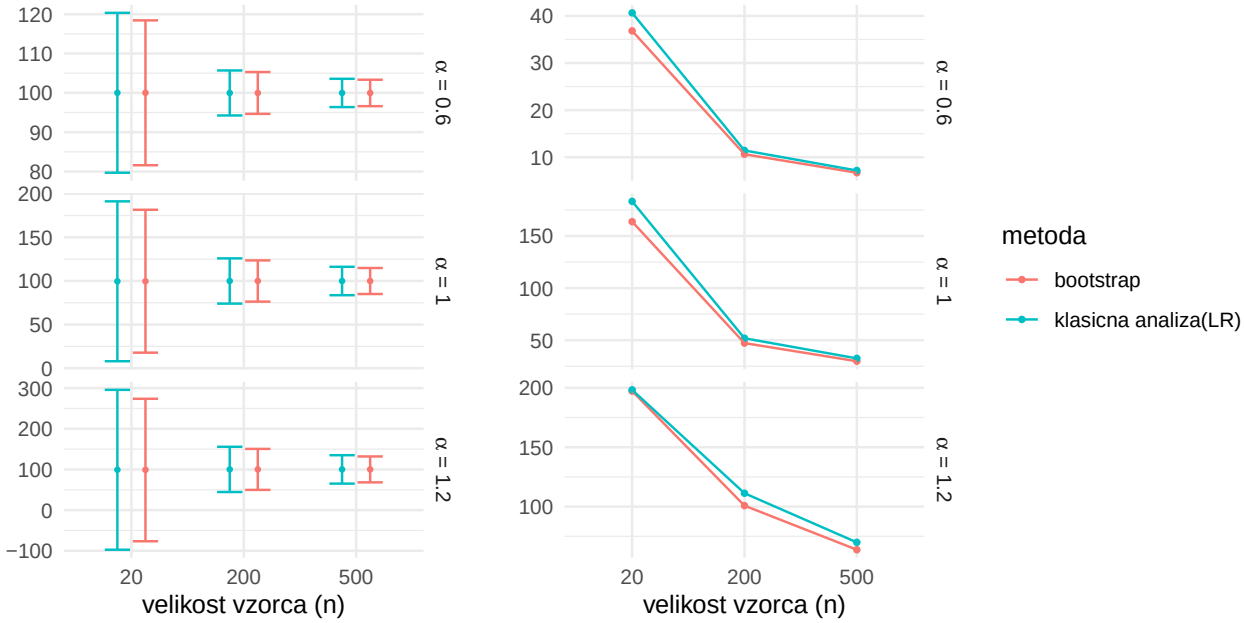
Za vsako kombinacijo faktorjev(*alpha* in *velikost vzorca*) sva generirala podatke, na katerih sva za vsako kombinacijo faktorjev izvedla linearno regresijo in poračunala intervale zaupanja za vse tri koeficiente(`Intercept`, `x1`, `x2`).

Enak postopek sva ponovila z metodo ponovnega vzorčenja bootstrap - naključno sva iz generiranih podatkov za vse kombinacije faktorjev izbrala podatke, na katerih sva potem izvedla linearno regresijo in izračunala intervale zaupanja za vse tri koeficiente (`Intercept`, `x1`, `x2`). Tak postopek ponovnega vzorčenja sva ponovila $m = 1000$, zato smo torej imeli 1000 bootstrap vzorcev.
Ves ta postopek sva ponovila 10000-krat in zaradi (predvsem) počasnosti *bootstrap* metode uporabila t.i. paralelno računanje(angl. *parallel computing*).

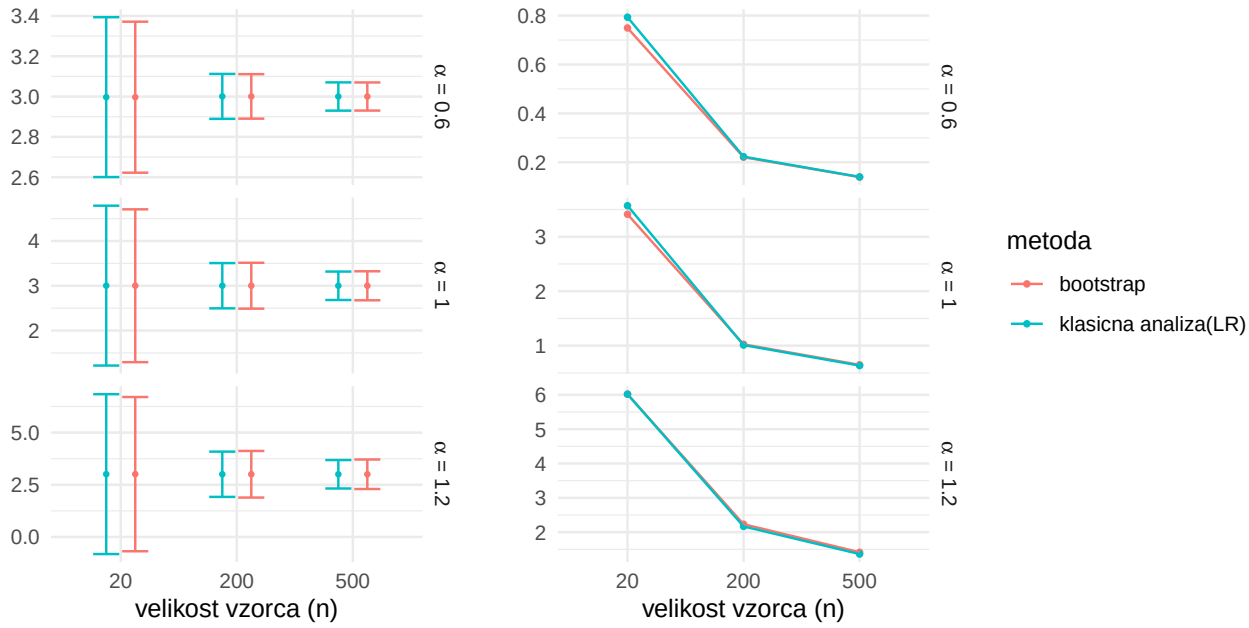
3.1 Analiza rezultatov podatkov

Jasno nam torej je, da se na rezultate, pridobljene s klasično metodo ne moremo ravno zanesti, saj kršenje predpostavk pri linearni regresiji močno vpliva na intervale zaupanja(tudi na ocene koeficientov). Pri obeh metodah pa seveda pri majhnem vzorcu($n = 20$) pričakujemo najširše intervale zaupanja, ki pa se potem z večanjem vzorca ožajo. Verjetno pa bodo intervali zaupanja pri obeh metodah(klasični test in bootstrap) približno enako široki.

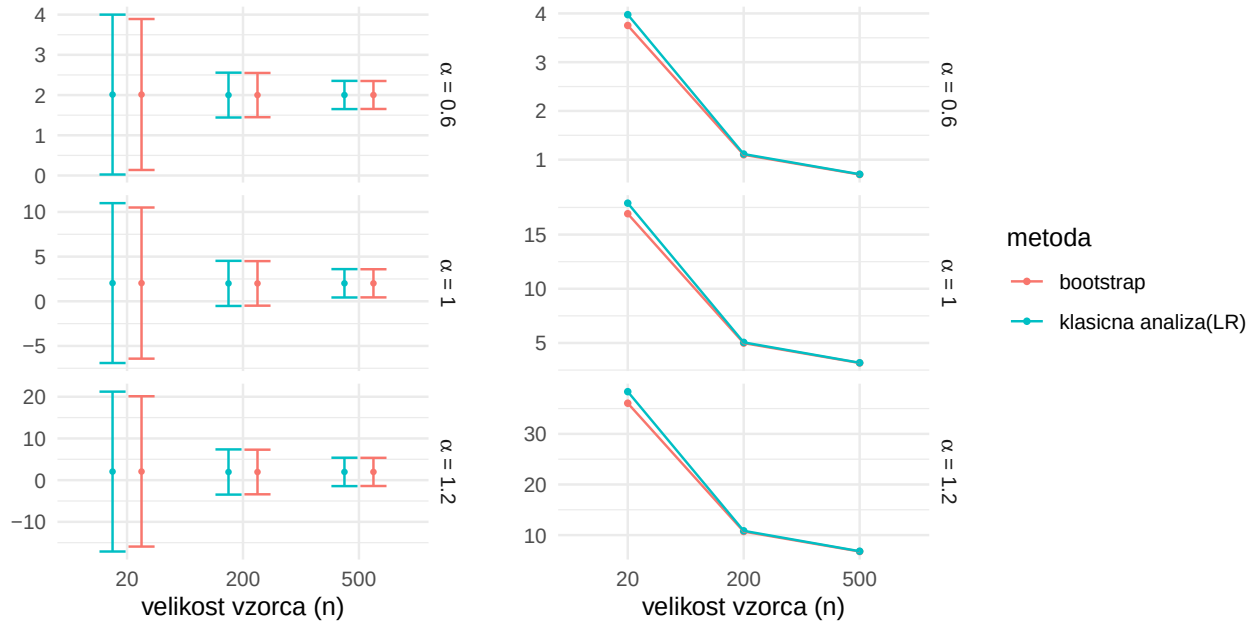
Na spodnjih grafih vidimo, da se intervali z večanjem vzorca res manjšajo, razlike med intervali s klasično analizo in bootstrapom pa so minimalne, največja razlika v širini intervala se opazi pri majhnem vzorcu($n = 20$). Težko pa rečemo, da faktor heteroskedastičnost(α) glede na metodo kako značilno vpliva na širino intervala.



Slika 3: Grafi intervalov zaupanja(levo) in širine intervalov zaupanja(desno) za prosti koeficient(Intercept).



Slika 4: Grafi intervalov zaupanja(levo) in širine intervalov zaupanja(desno) za koeficient pri x1(beta1).



Slika 5: Grafi intervalov zaupanja(levo) in širine intervalov zaupanja(desno) za koeficient pri x2(beta2).

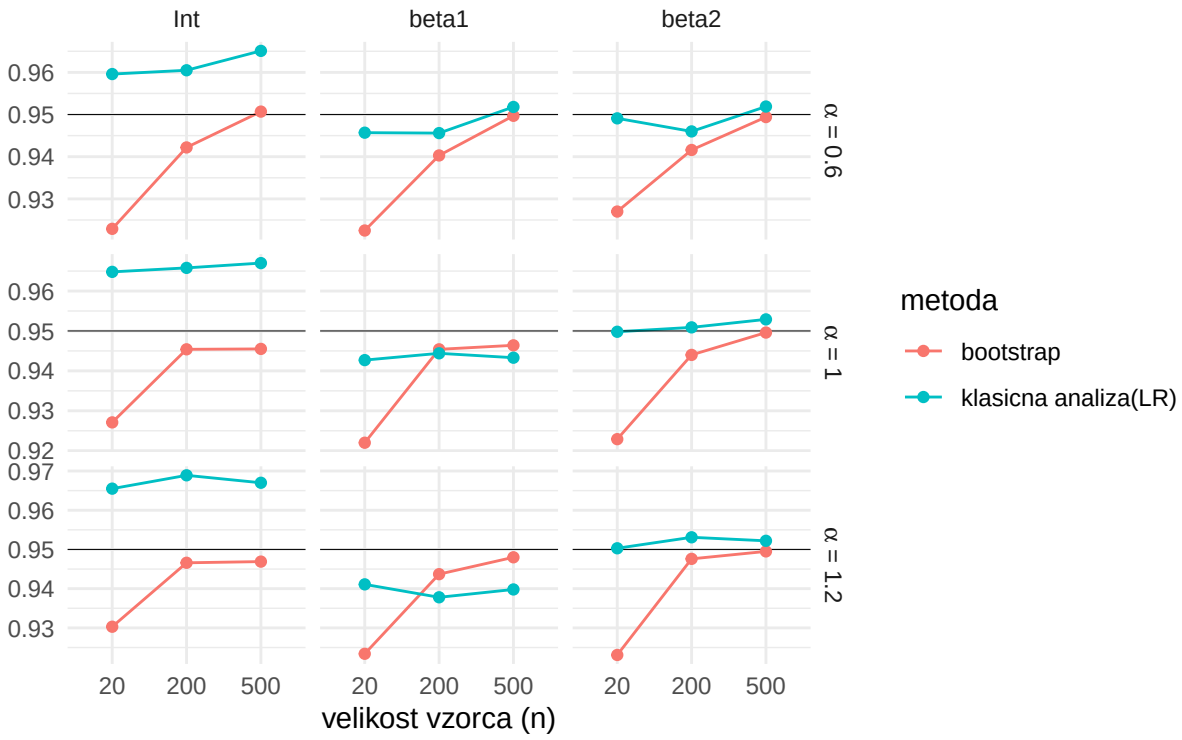
3.1.1 Pokritost

Pokritost se nanaša na odstotek primerov, ko interval zaupanja dejansko zajame resnično vrednost parametra. Mi imamo 95% interval zaupanja, torej pričakujemo, da interval zajame resnično vrednost v 95% ponovljenih vzorčenj. Rezultate si lahko ogledamo v spodni tabeli in grafu.

Tabela 1: Pokritost pri posameznih koeficientih in kombinacijah alphe in velikosti vzorca(n).

		intercept		beta1		beta2	
alpha	velikost vzorca	klasicna	bootstrap	klasicna	bootstrap	klasicna	bootstrap
0.6	20	0.9596	0.9229	0.9457	0.9225	0.9491	0.9270
0.6	200	0.9605	0.9422	0.9456	0.9403	0.9460	0.9416
0.6	500	0.9651	0.9507	0.9518	0.9497	0.9519	0.9494
1.0	20	0.9648	0.9271	0.9427	0.9220	0.9498	0.9229
1.0	200	0.9658	0.9454	0.9444	0.9454	0.9509	0.9440
1.0	500	0.9670	0.9455	0.9433	0.9464	0.9529	0.9496
1.2	20	0.9655	0.9303	0.9411	0.9234	0.9503	0.9231
1.2	200	0.9689	0.9466	0.9378	0.9437	0.9531	0.9476
1.2	500	0.9670	0.9469	0.9398	0.9480	0.9522	0.9495

Vidimo, da se pokritost razlikuje glede na metodo, velikost vzorca in faktor α. Vidimo, da je v večini primerov pokritost pri klasičnem testu(linearna regresija) boljše, ampak težko ocenimo, če je to res bolje, saj so kršene predpostavke in ocene koeficientov in intervali zaupanja takrat niso zanesljivi. Največje razlike v pokritosti so pri prostem členu(`Intercept`) in pri majhnem vzorcu $n = 20$.



Slika 6: Graf pokritosti za vse tri koeficiente.

4 Zaključek

Pri primerjavi intervalov zaupanja "klasične" metode (*linearne regresije*), ko so predpostavke te kršene, in metode ponovnega vzorčenja (*bootstrap*), nisva opazila drastičnih razlik. Intervali zaupanja so bili pri metodi ponovnega vzorčenja nekoliko ožji, največja razlika v širini intervala pa se opazi pri majhnem vzorcu ($n = 20$).

Pri pokritosti se izkaže, da ima klasična analiza boljšo pokritost že pri majhnem vzorcu, *bootstrap* metoda pa se približa 95% pokritosti šele pri zelo velikem vzorcu $n = 500$.

Glede na analizo bi težko rekla, da je ena metoda veliko boljša od druge. Vendar pa bi vseeno ob kršenju predpostavk uporabila metodo ponovnega vzorčenja, saj se je ta izkazala za malenkost boljšo in z velikim številom vzorcev (vzorčenja) odstranimo ekstremne robne vrednosti.