

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: *obstoje 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

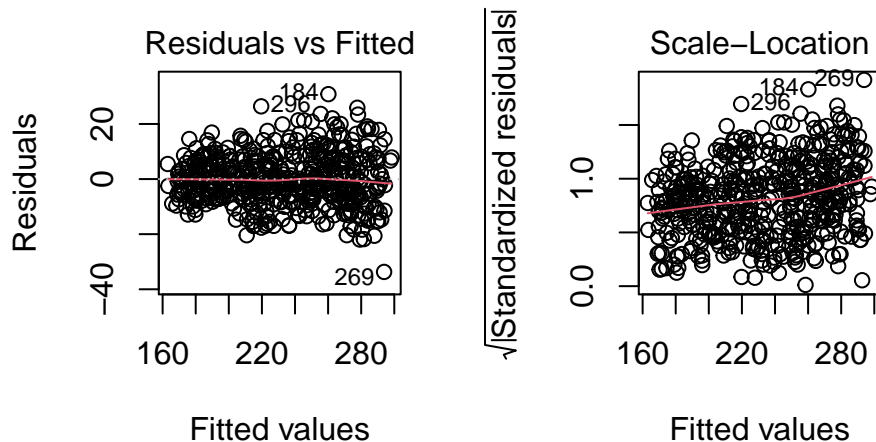
$$y_i = 100 + 3 \cdot x_{1i} + 2 \cdot x_{2i} + \epsilon_i.$$

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 $\alpha \in \{0.6, 1, 1.2\}$.

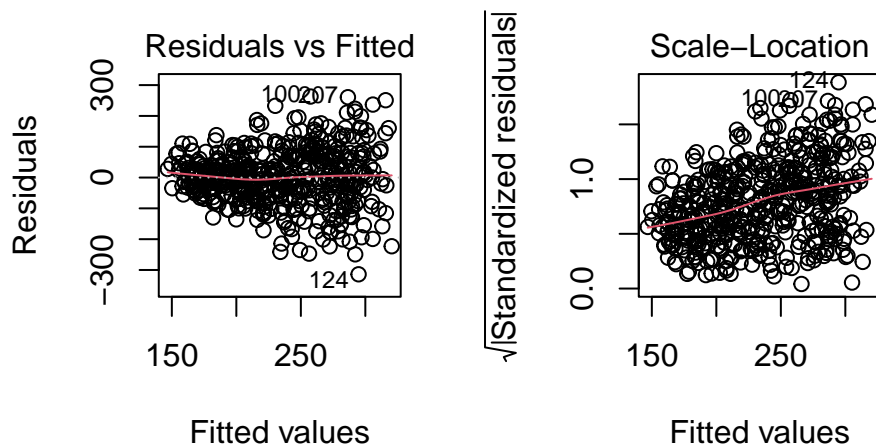
Velikost vzorca pa enak $n = (20, 200, 500)$, ker naju zanima kako se bo bootstrap metoda obnesla tudi na zelo majhnem vzorcu - s tem je tudi kršena predpostavka o dovolj velikem vzorcu pri linearni regresiji.

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 parametrih $\alpha=0,6$.



Slika 2: Grafi ostankov pri parametrih $\alpha=1,2$.

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(α 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**, x_1 , x_2).

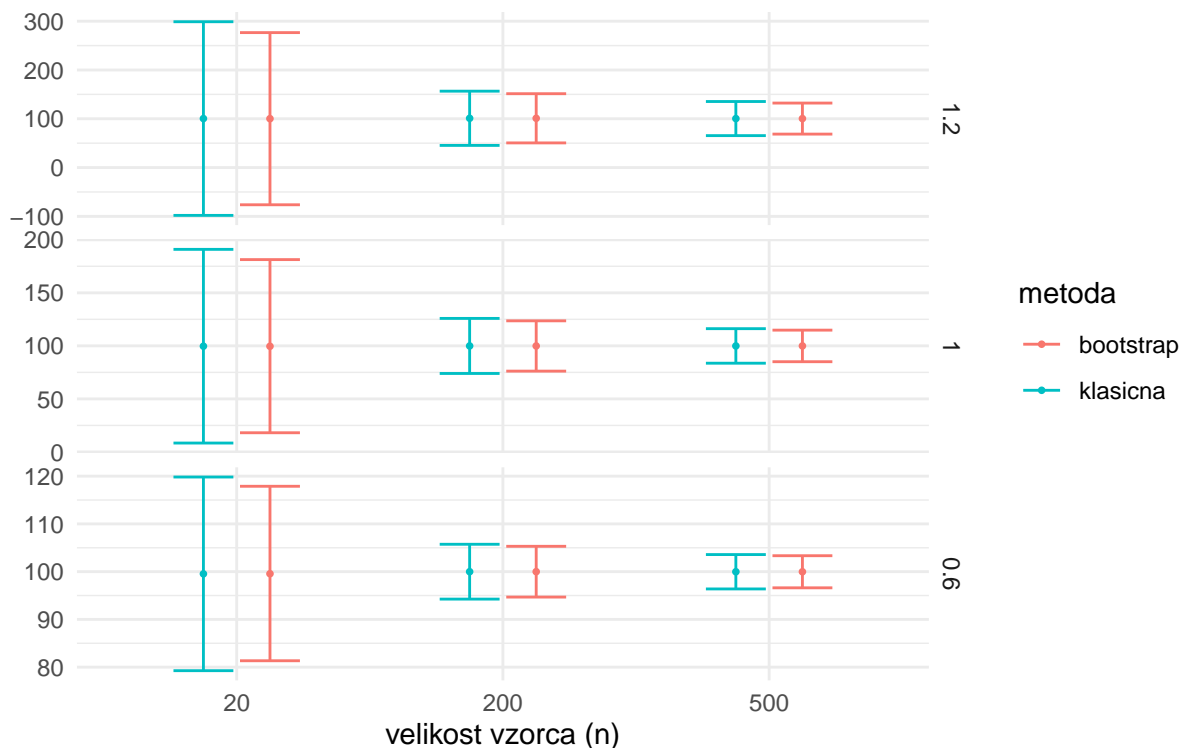
Enak postopek sva ponovila z metodo ponovnega vzorčenja bootstrap - naključno sva iz generiranih podatkov za vse kombinacije faktorjev izbrala podatke, na katerih sva potem izvedla linearno regresijo in izračunala intervale zaupanja za vse tri koeficiente (**Intercept**, **x1**, **x2**). Tak postopek ponovnega vzorčenja sva ponovila $m = 1000$, zato smo torej imeli 1000 bootstrap vzorcev.

Ves ta postopek sva ponovila 1000-krat in zaradi (predvsem) počasnosti *bootstrap* metode uporabila in uporabila t.i. paralelno računanje(angl. *parallel computing*).

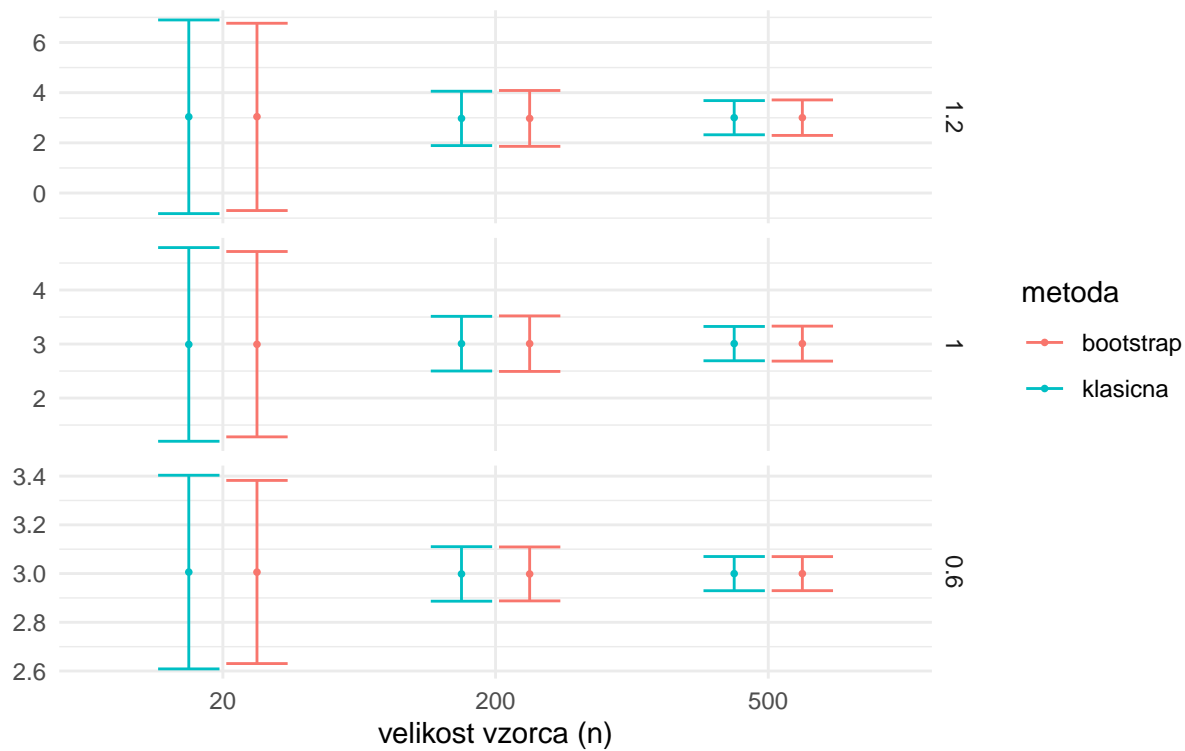
3.1 Analiza rezultatov podatkov

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

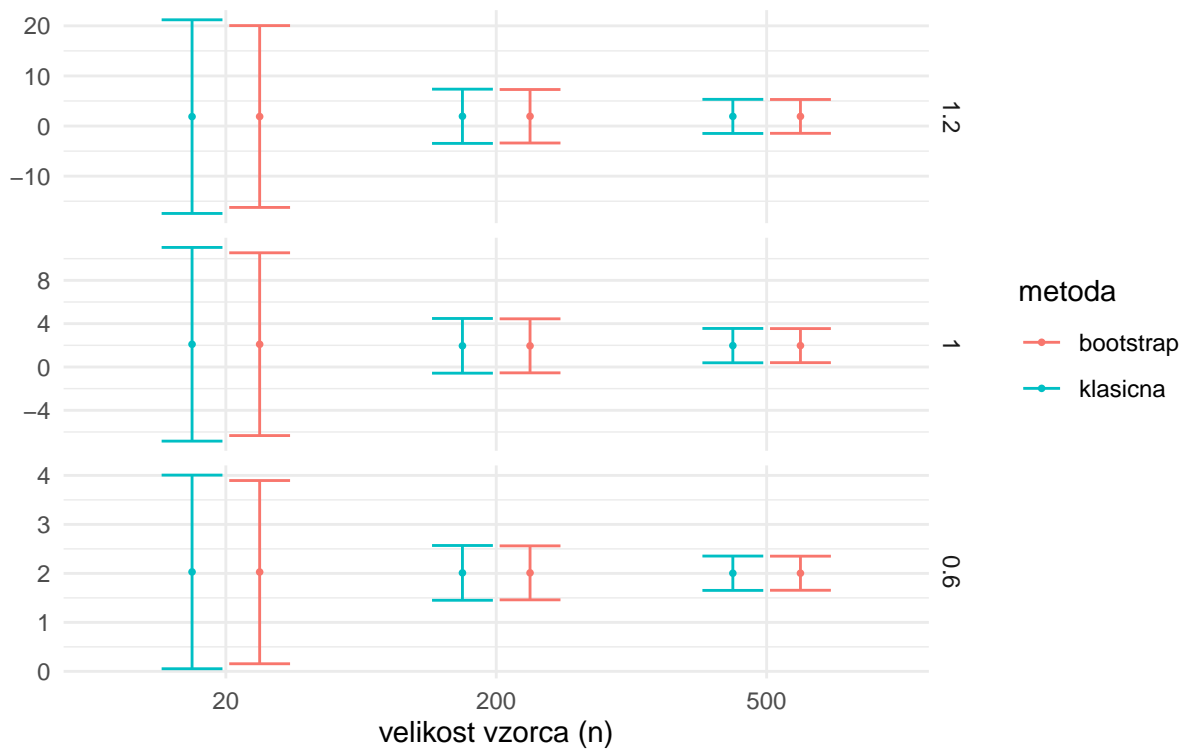
Na spodnjih grafih vidimo, da se intervali z večanjem vzorca res manjšajo, razlike med intervali s klasično analizo in bootstrapom pa so minimalne, največja razlika v širini intervala se opazi pri majhnem vzorcu($n = 20$).



Slika 3: Grafi intervalov zaupanja za prosti koeficient(Intercept).



Slika 4: Grafi intervalov zaupanja za koeficient pri $x_1(\beta_1)$.



Slika 5: Grafi intervalov zaupanja za koeficient pri $x_2(\beta_2)$.

alpha	velikost vzorca	intercept		beta1		beta2	
		klasicna	bootstrap	klasicna	bootstrap	klasicna	bootstrap
0.6	20	0.9610	0.9297	0.9437	0.9243	0.9470	0.9277
0.6	200	0.9630	0.9407	0.9377	0.9373	0.9460	0.9457
0.6	500	0.9637	0.9457	0.9477	0.9523	0.9453	0.9393
1.0	20	0.9610	0.9297	0.9437	0.9243	0.9470	0.9277
1.0	200	0.9630	0.9407	0.9377	0.9373	0.9460	0.9457
1.0	500	0.9637	0.9457	0.9477	0.9523	0.9453	0.9393
1.2	20	0.9610	0.9297	0.9437	0.9243	0.9470	0.9277
1.2	200	0.9630	0.9407	0.9377	0.9373	0.9460	0.9457
1.2	500	0.9637	0.9457	0.9477	0.9523	0.9453	0.9393