Domača naloga 3

Neža Kržan, Tom Rupnik Medjedovič

1 Cilji naloge

Želiva preučiti uporabo metode ponovnega vzorčenja, v primeru ko klasičnimi testom ne moremo popolno zaupati zaradi kršenja predpostavk. 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 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) in zaradi tega bova s pomočjo testa `boxCox` izvedla primerno transformacijo odzivne spremenljivke (ta bo v vseh primerih `log` transformacija). Na koncu bova med seboj primerjala intervale zaupanja dobljene z linearno regresijo (`lm`) in metodo ponovnega vzorčenja(bootstrap in permutacijski test). Primerjavo bova naredila tako na rezultatih pred in po transformaciji, vendar pa moramo paziti, saj rezultati pred in po transformaciji med seboj niso primerljivi.

Pričakujeva, da bomo z metodo ponovnega vzorčenja dobili boljše rezultate.

2 Generiranje podatkov

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

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

pri čemer:

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

torej

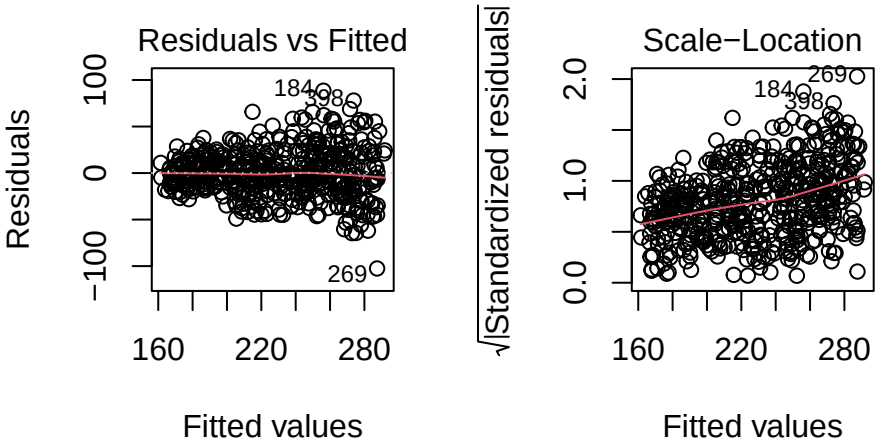
$$y_i = 100 + 3 \cdot x_{1i} + 2 \cdot x_{2i}^{\gamma} + \epsilon_i.$$

Kot sva že omenila bova spreminjala faktorja γ in α. S faktorjem γ bomo kršili prespostavko o linearni zvezi, saj bo ta zavzel vrednosti 0.8 in 1.4. S faktorjem α pa bomo kršili predpostavko konstantne variance napak, saj se ta z večanjem vrednosti x_1 povečuje. Ta zavzame vrednosti $\alpha \in \{0.6, 1, 1.2\}$.

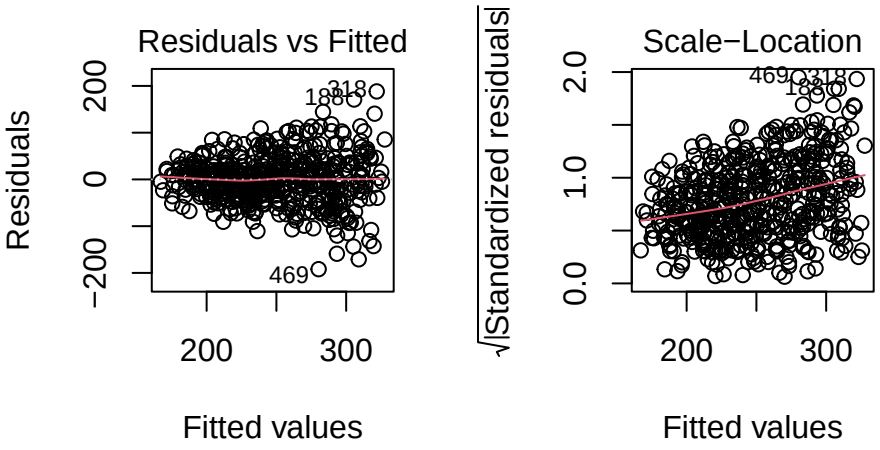
Velikost vzorca pa enak $n = 500$, saj je to dovolj velik vzorec za pravilno delovanje linearne regresije.

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 in gamma=0,8.



Slika 2: Grafi ostankov pri paramatrih alpha=1,2 in gamma=1,4.

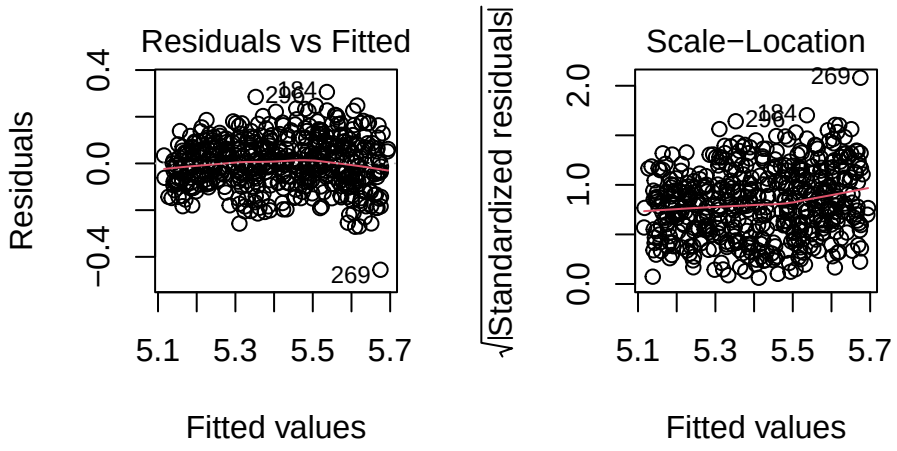
V obeh primerih lahko na desnem grafu opazimo, da se z večanjem vrednosti povečuje tudi variabilnost napak (naraščajoč trend). Na levem grafu lahko sorazmerno s faktorjem γ pričakujemo odstopanje vrednosti na robovih. V primeru $\gamma = 0.8$ opazimo, da so na robovih nekoliko nižje vrednosti, v primeru $\gamma = 1.4$ pa nekoliko višje vrednosti.

Kot sva že napisala zgoraj, bova zaradi kršenja dveh predpostavk(linerana odvisnost in homoskedastičnost) podatke ustrezno transformirala. Za tako kršene predpostavke ponavadi uporabljamo logaritemske transformacije podatkov.

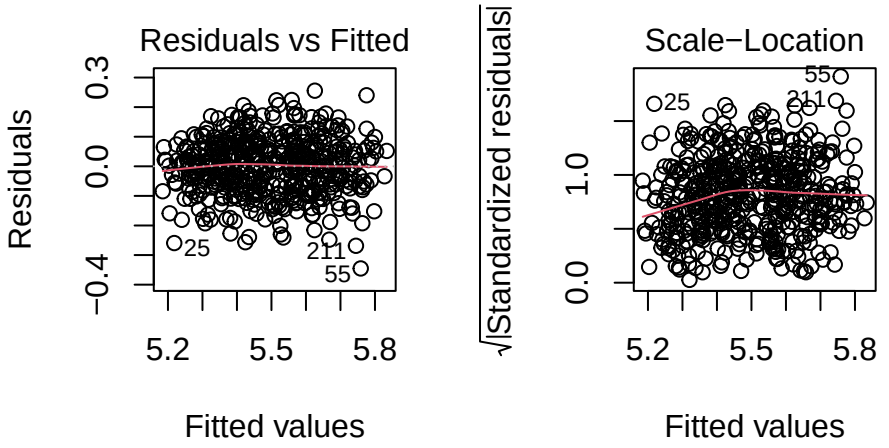
alpha	gamma	spodnja meja IZ	zgornja meja IZ
0.6	0.8	-0.623	0.030
0.6	1.4	-0.342	0.302
1.0	0.8	-0.143	0.429
1.0	1.4	-0.137	0.444
1.2	0.8	-0.012	0.441
1.2	1.4	0.026	0.524

Za izbiro primerne transformacije si bova pomagala s `boxCox` testom - za vsako kombinacijo parametrov preverimo ali se vrednost $\lambda = 0$ (`log` transformacija) nahaja znotraj 95% intervala optimalnega parametra λ, ki ga vrne funkija `powerTransform`. V spodnji tabeli lahko vidimo, da je $\lambda = 0$ res vsebovana v vseh 95% intervalih zaupanja, razen v zadnjem primeru, torej je primerna transformacija podatkov logaritemska. Pri zadnjem primeru, pa je spodnja meja 95% intervala zaupanja tako blizu vrendosti 0, da prav tako lahko uporabimo logaritemsko transformacijo, ker bomo dobili boljše rezultate in bodo predopstavke bolje izpolnjene.

Če si sedaj ponovno pogledamo grafe ostankov transformiranih podatkov z istimi kombinacijami faktorjev kot na zgornjih grafih ostankov, vidimo, da so ostanki na grafih razpršeni naključno, torej predpostavki(linearna odvisnost in homoskedastičnost) nista kršeni.



Slika 3: Grafi ostankov transformiranih podatkov pri paramatrih alpha=0,6 in gamma=0,8.



Slika 4: Grafi ostankov transformiranih podatkov pri paramatrih alpha=0,6 in gamma=1,4.

3 Metoda ponovnega vzorčenja

Pri ponovnem vzorčenju bova uporabila metodo bootstrap, pri kateri bo število bootstrap vzorcev enako $m = 1000$. Za vsako kombinacijo faktorjev(*alpha*, *gamma* in *velikost vzorca*) sva generirala podatke, na katerih sva za vsako kombinacijo faktorjev torej 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.