

Bayesova statistika

Seminarska naloga

Sara Bizjak | 27202020

Februar 2022

Predstavitev problema

V sklopu seminarske naloge bomo pri multipli linearni regresiji preko simulacij primerjali Bayesov in frekventistični pristop.

Pri obeh pristopih uporabimo običajni normalni model s pomočjo funkcij `bayesx` in `lm`. Primerjavo izvedemo glede na tri različne scenarije:

1. normalno porazdeljena napaka z majhno varianco: $\epsilon_1 \sim N(0, 2^2)$,
2. normalno porazdeljena napaka z veliko varianco: $\epsilon_2 \sim N(0, 100^2)$,
3. asimetrično porazdeljena napaka (hi-kvadrat): $\epsilon_3 \sim \chi_{df=1}^2$.

Za velikost vzorca vzamemo $n = 100$, število ponovitev simulacij pa naj bo 1000. Generiramo 5 različnih pojasnjevalnih spremenljivk, za katere izberemo različno močne efekte in jih simuliramo neodvisno eno od druge.

$$\begin{aligned}X_1 &\sim N(10, 4) \\X_2 &\sim Ber(0.8) \\X_3 &\sim N(15, 25) \\X_4 &\sim Bin(10, 0.7) \\X_5 &\sim Ber(0.2)\end{aligned}$$

Spremenljivko Y generiramo kot

$$Y = intercept + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + napaka,$$

kjer si za koeficiente izberemo

- $intercept = 25$
- $\beta_1 = 0$

- $\beta_2 = 5$
- $\beta_3 = 50$
- $\beta_4 = 4$
- $\beta_5 = 10$

Tukaj je *intercept* začetna vrednost, β_1 predstavlja koeficient brez učinka, β_3 koeficient z močnim vplivom, β_4 koeficient s srednje močnim vplivom, β_5 pa koeficient s šibkim vplivom.

Analiza konvergencije

Za vsak posamezen scenarij (izmed zgoraj opisanih treh možnih) preverimo konvergenco pri eni simulaciji. Preverimo MCMC verige za vsak koeficient in tako preverimo, ali je potreba po dodatnem *burn-inu* (v funkciji je že vgrajen burn-in velikosti 2000 in *thinning* velikosti 10). Pogledamo si tudi avtokorelacije in podvzorce (podobno kot pri zadnji domači nalogi, ko smo opazovali konvergenco). Prikažimo grafe za $\beta_1, \beta_2, \beta_3, \beta_4$ in β_5 .

Model z majhno varianco

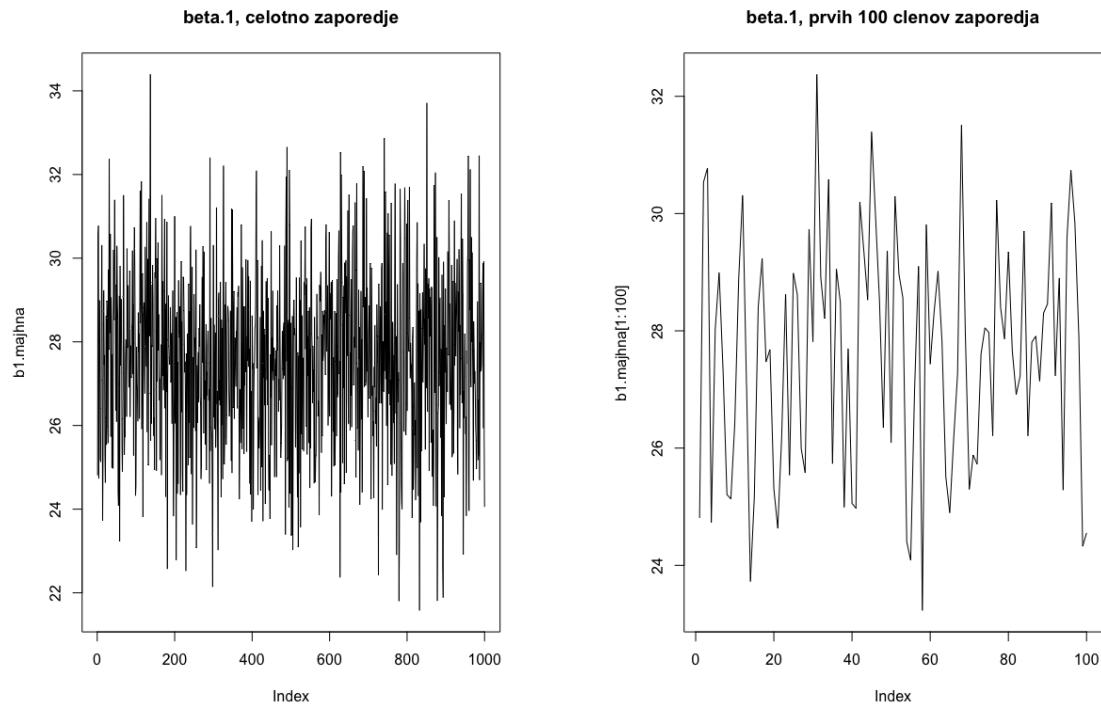


Figure 1: Grafa s prikazom konvergencije za koeficient β_1 .

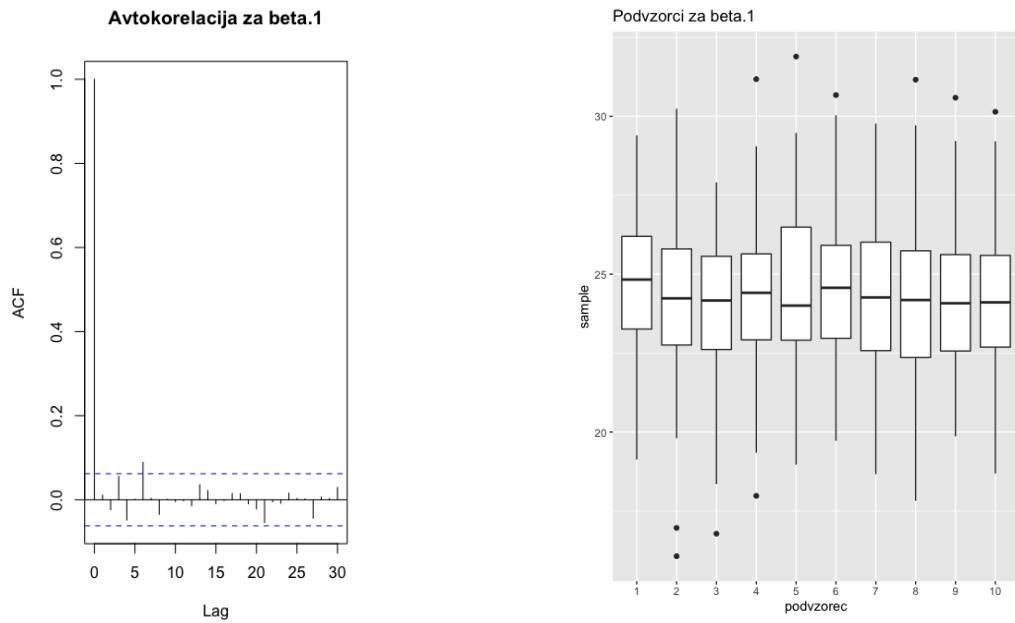


Figure 2: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_1 .

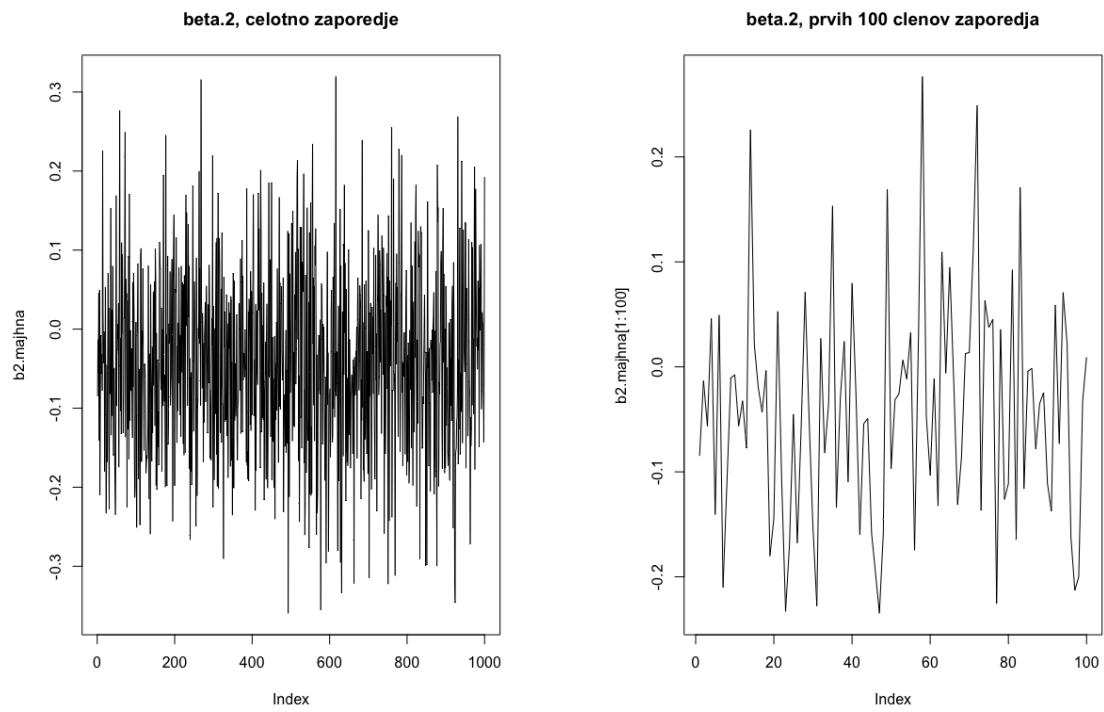


Figure 3: Grafa s prikazom konvergencije za koeficient β_2 .

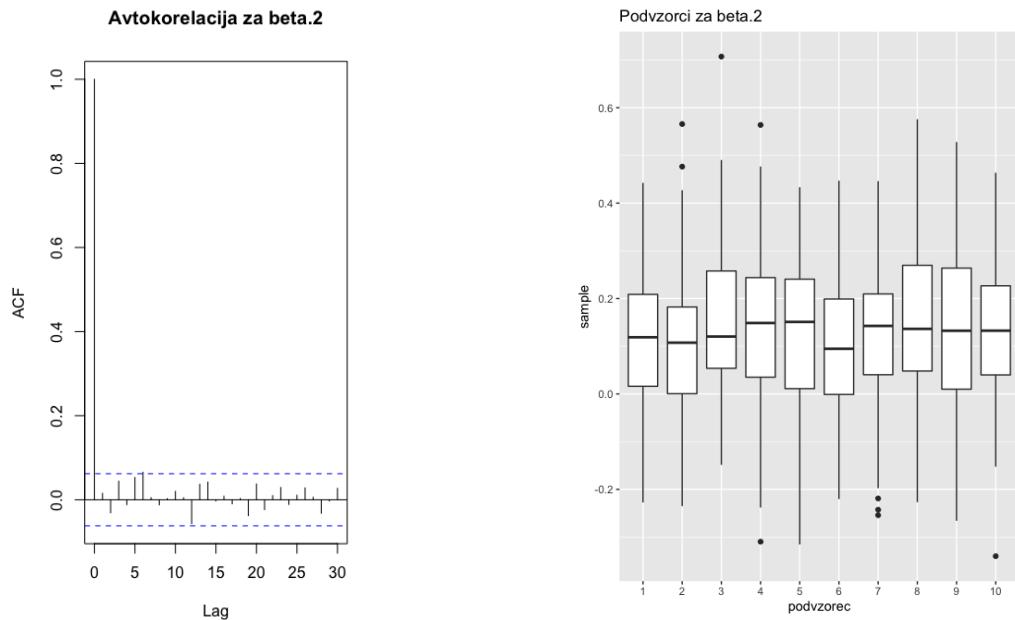


Figure 4: Grafa s prikazom avtokorelacij in podvzorcev za koeficient β_2 .

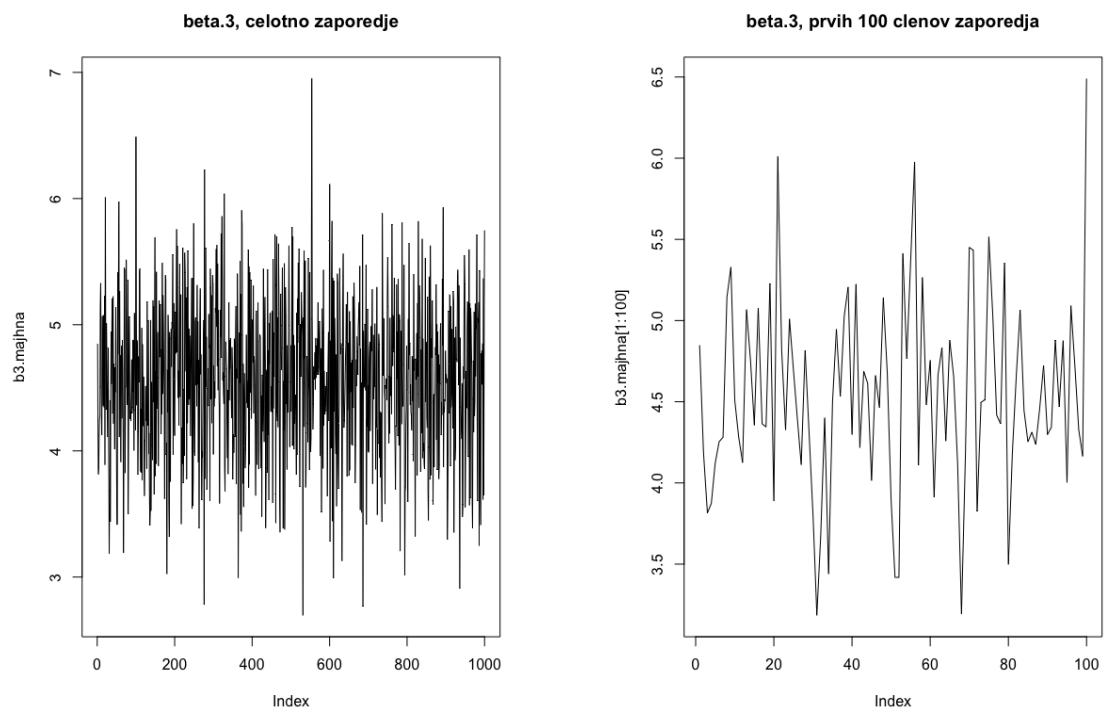


Figure 5: Grafa s prikazom konvergencije za koeficient β_3 .

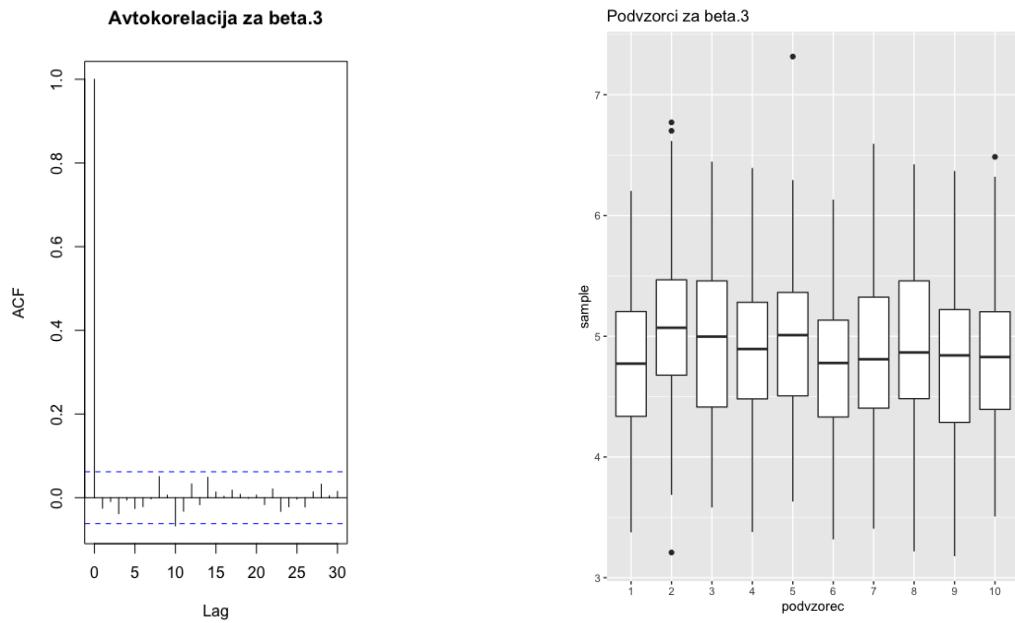


Figure 6: Grafa s prikazom avtokorelacij in podvzorcev za koeficient β_3 .

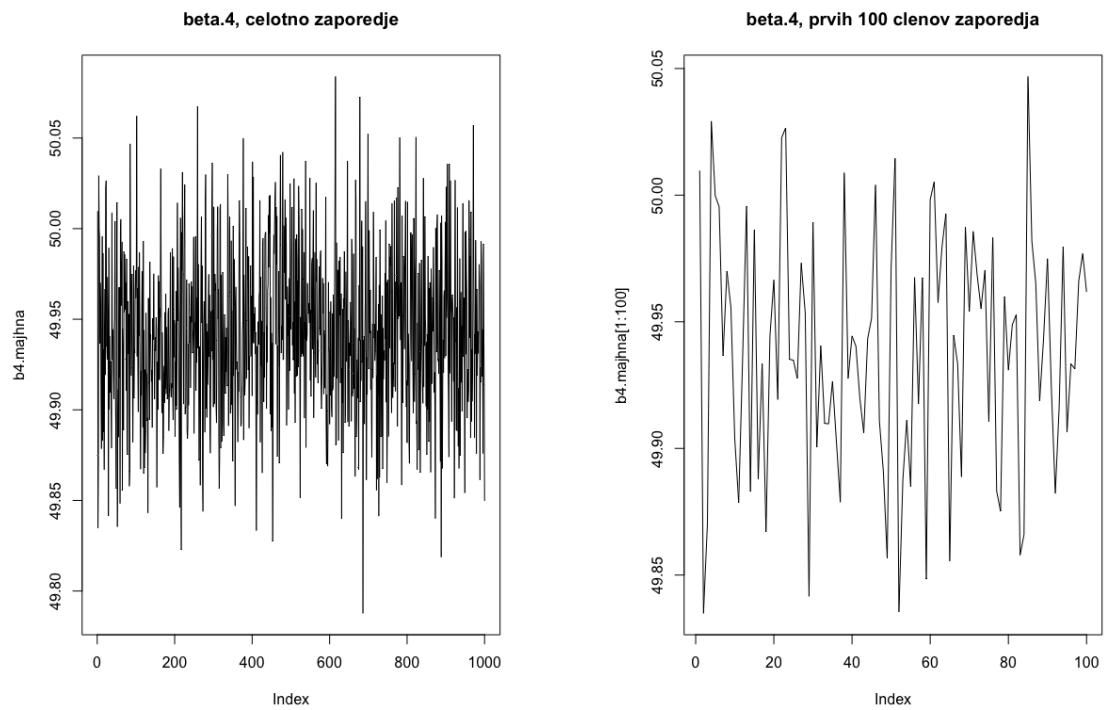


Figure 7: Grafa s prikazom konvergencije za koeficient β_4 .

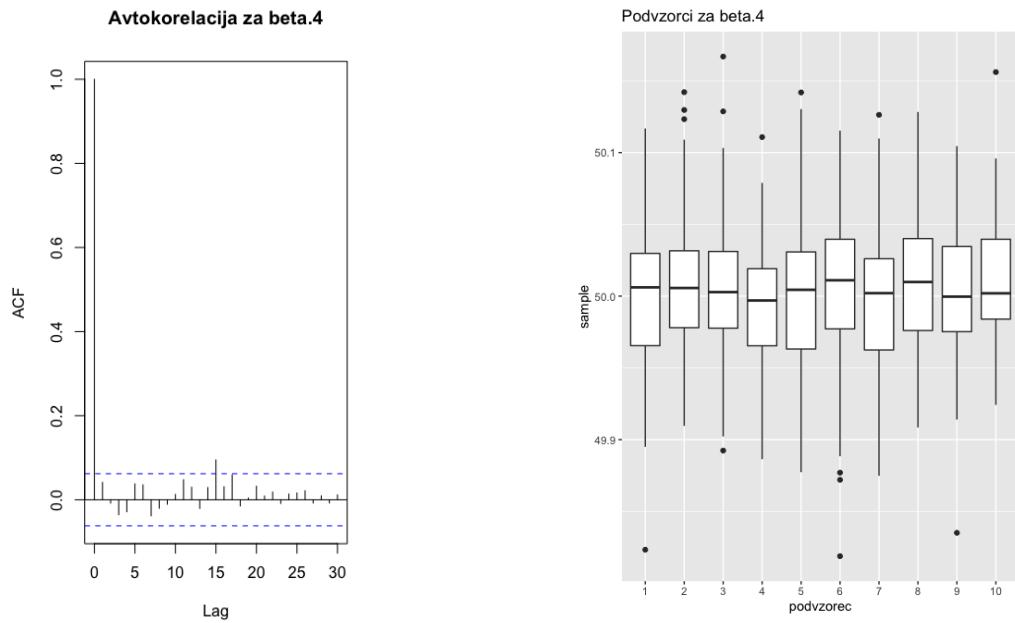


Figure 8: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_4 .

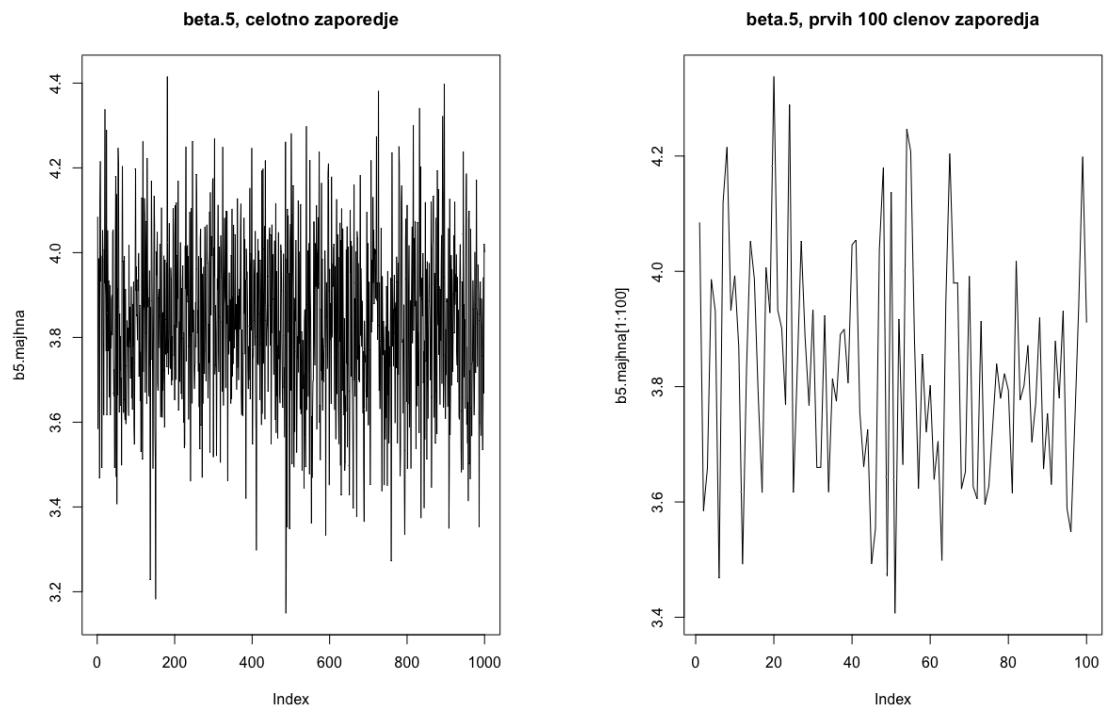


Figure 9: Grafa s prikazom konvergencije za koeficient β_5 .

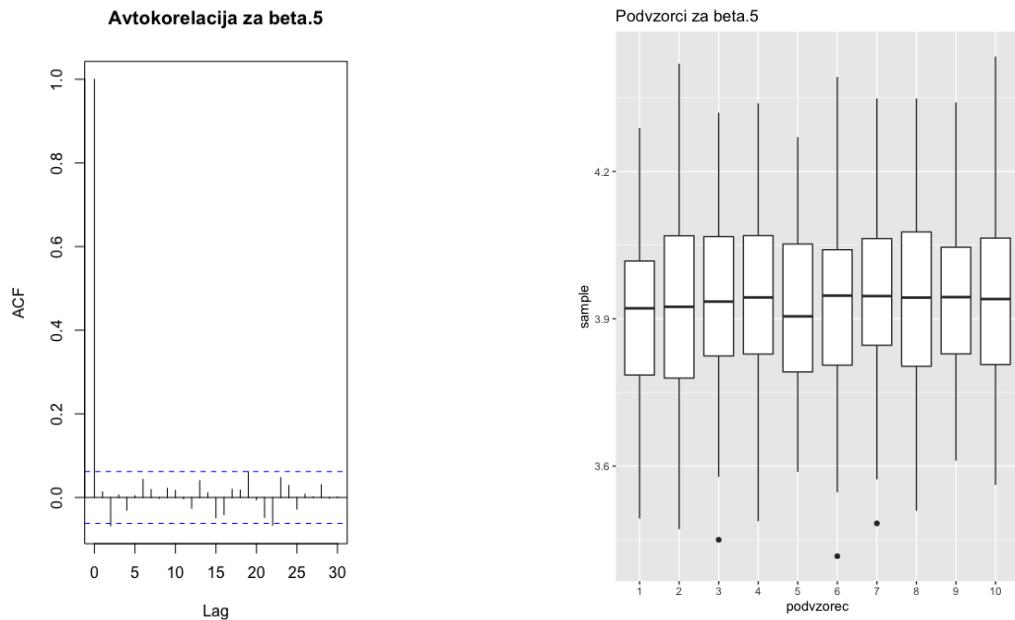


Figure 10: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_5 .

Model z veliko varianco

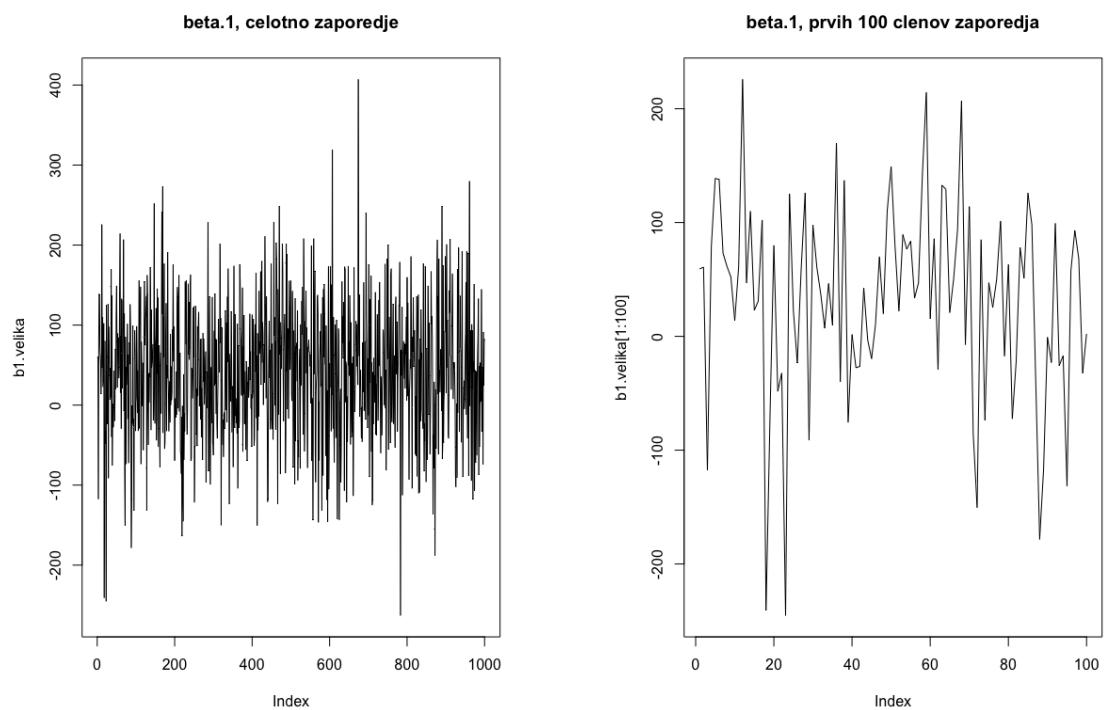


Figure 11: Grafa s prikazom konvergencije za koeficient β_1 .

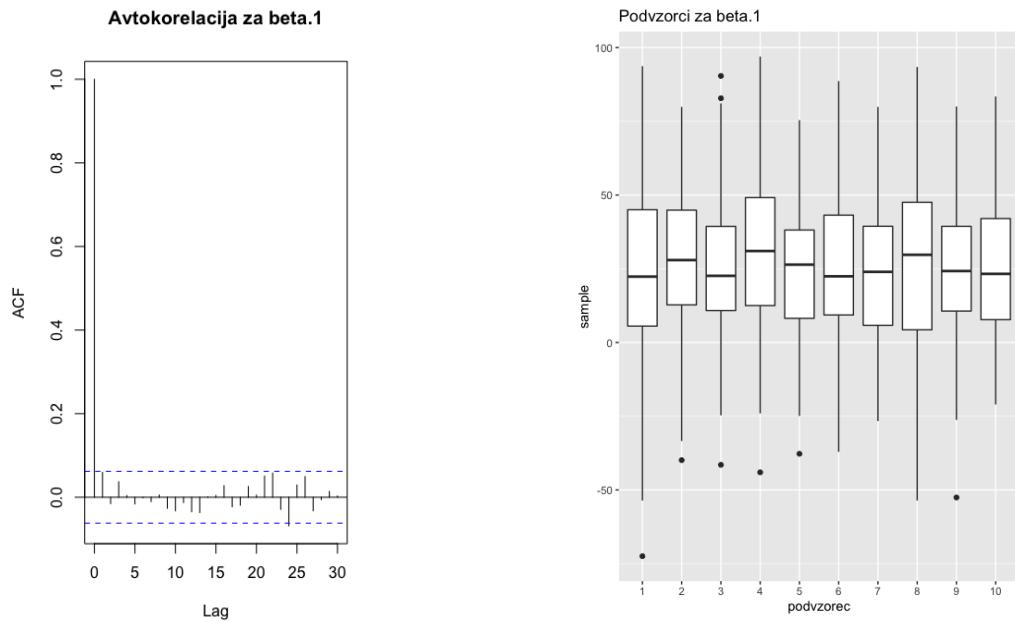


Figure 12: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_1 .

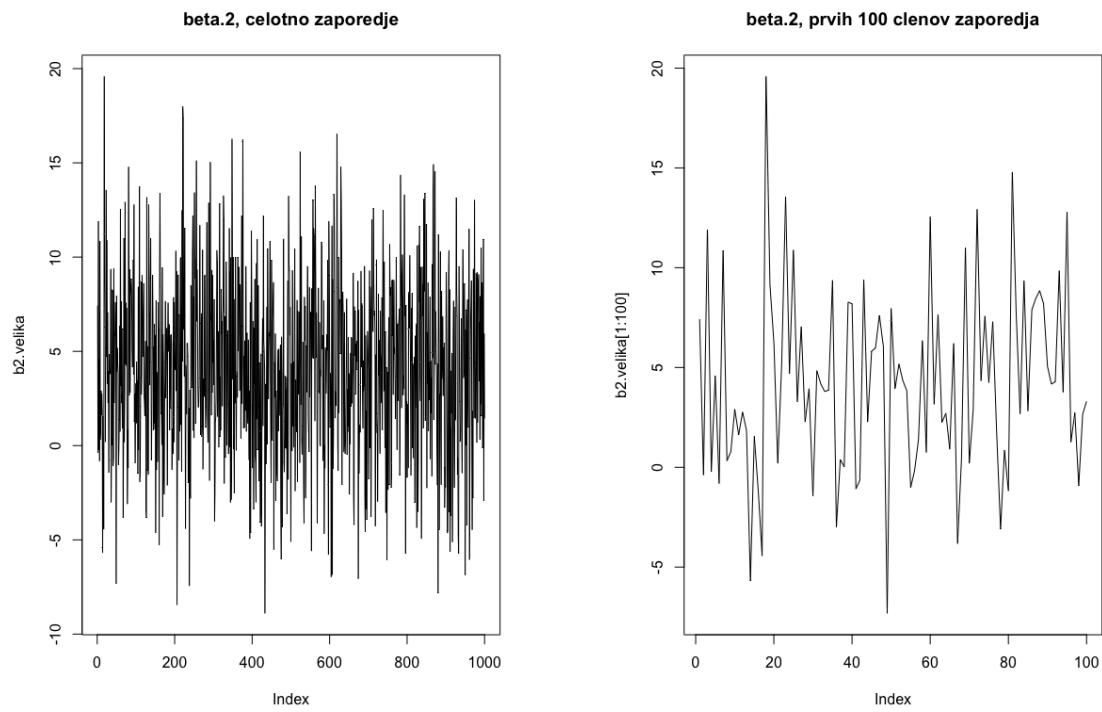


Figure 13: Grafa s prikazom konvergencije za koeficient β_2 .

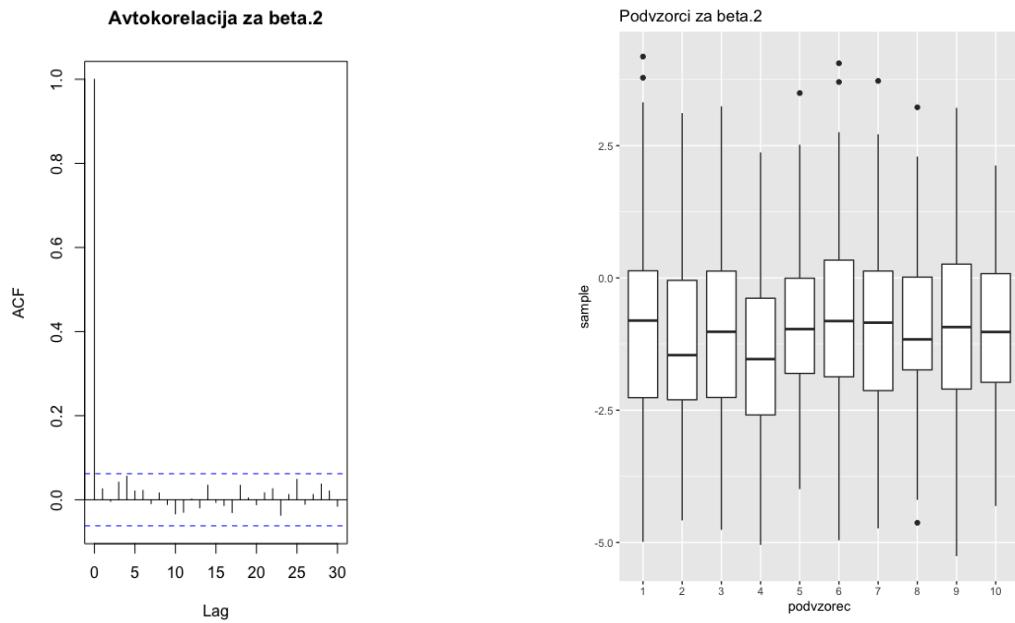


Figure 14: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_2 .

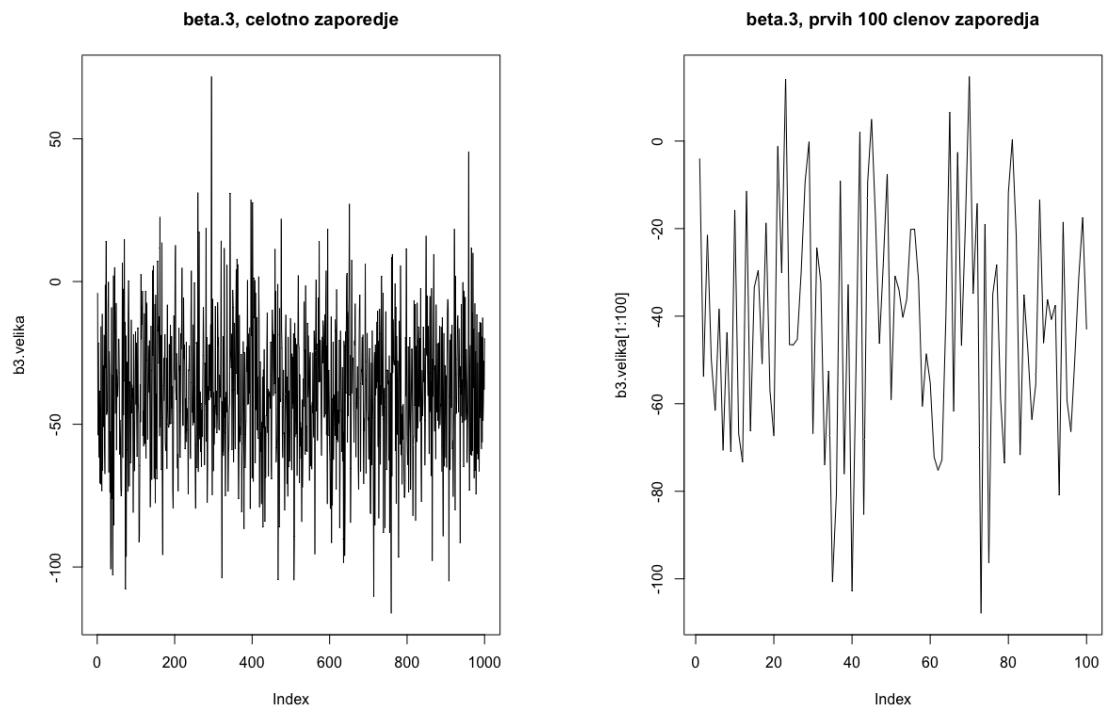


Figure 15: Grafa s prikazom konvergencije za koeficient β_3 .

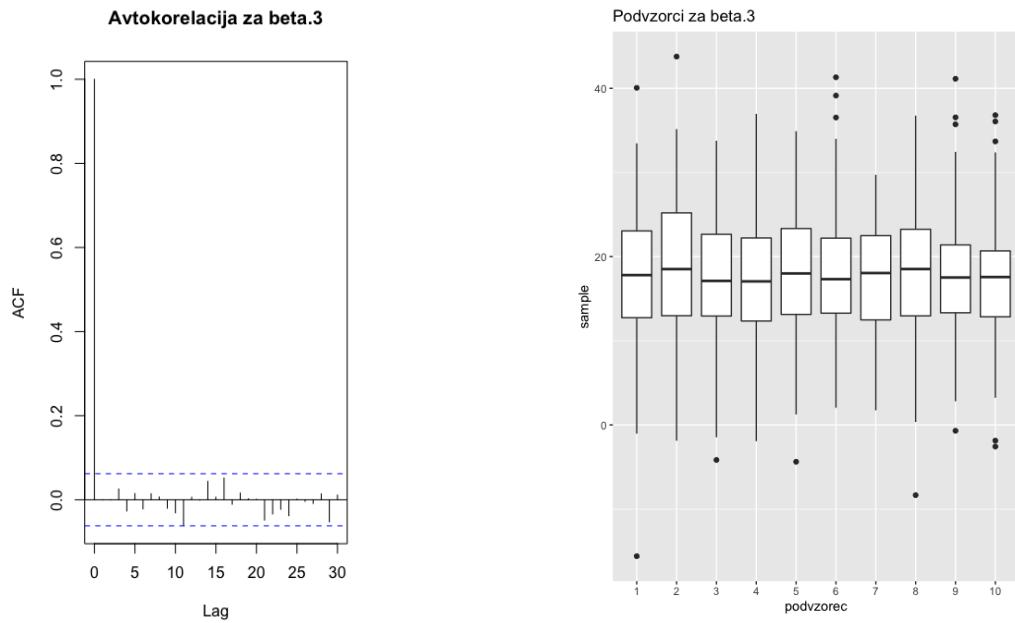


Figure 16: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_3 .

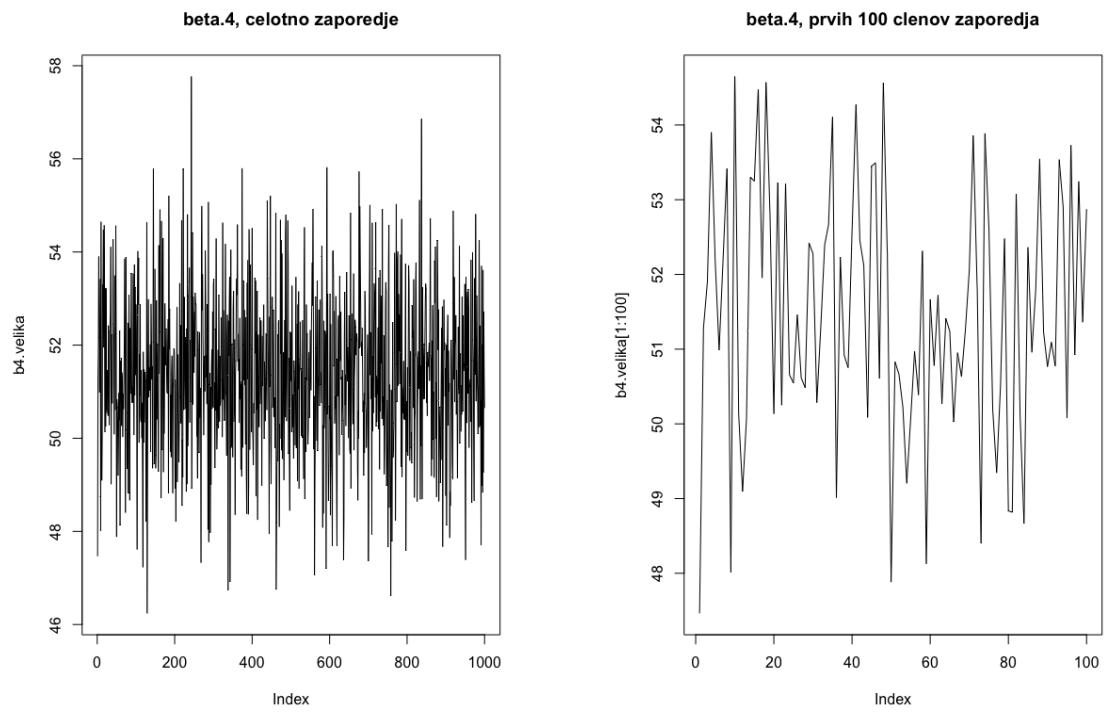


Figure 17: Grafa s prikazom konvergencije za koeficient β_4 .

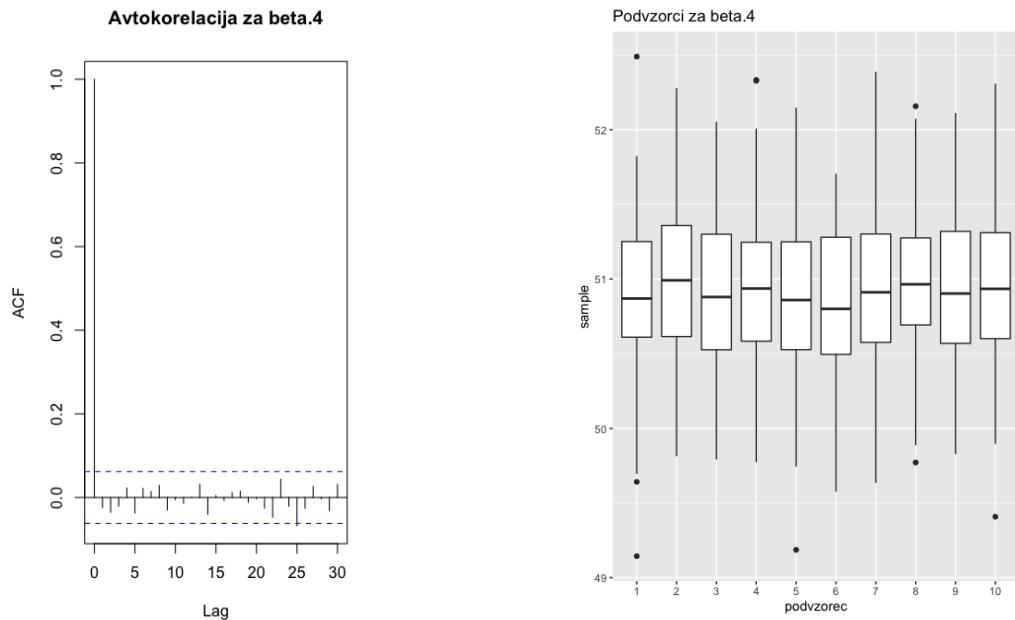


Figure 18: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_4 .

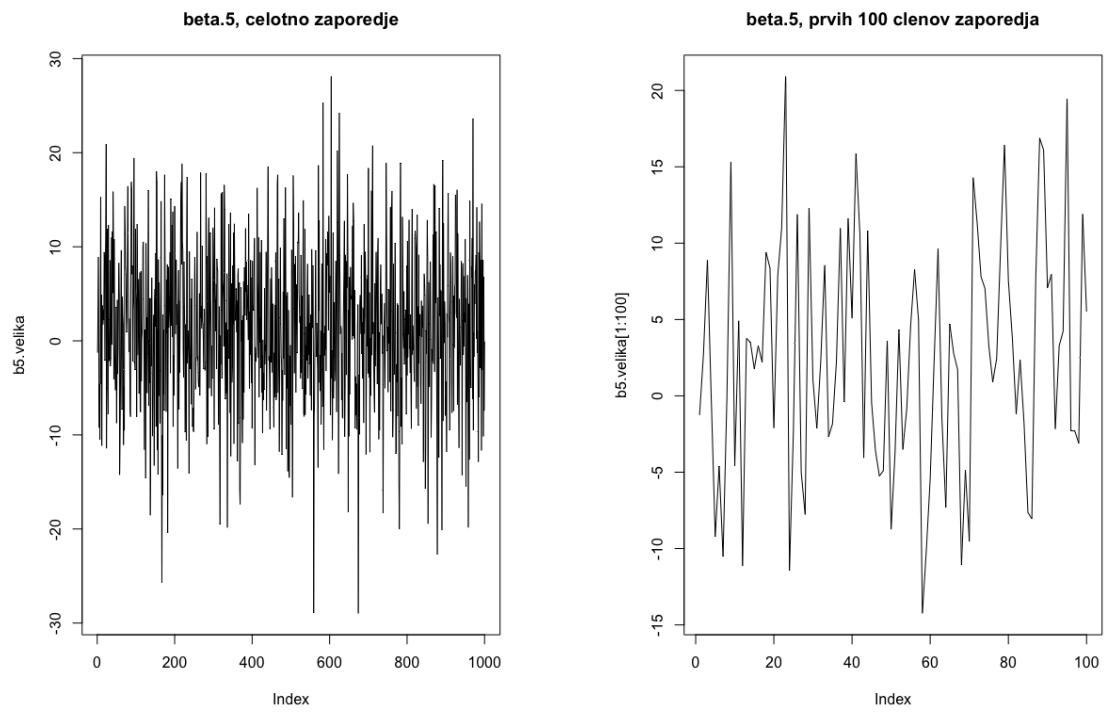


Figure 19: Grafa s prikazom konvergencije za koeficient β_5 .

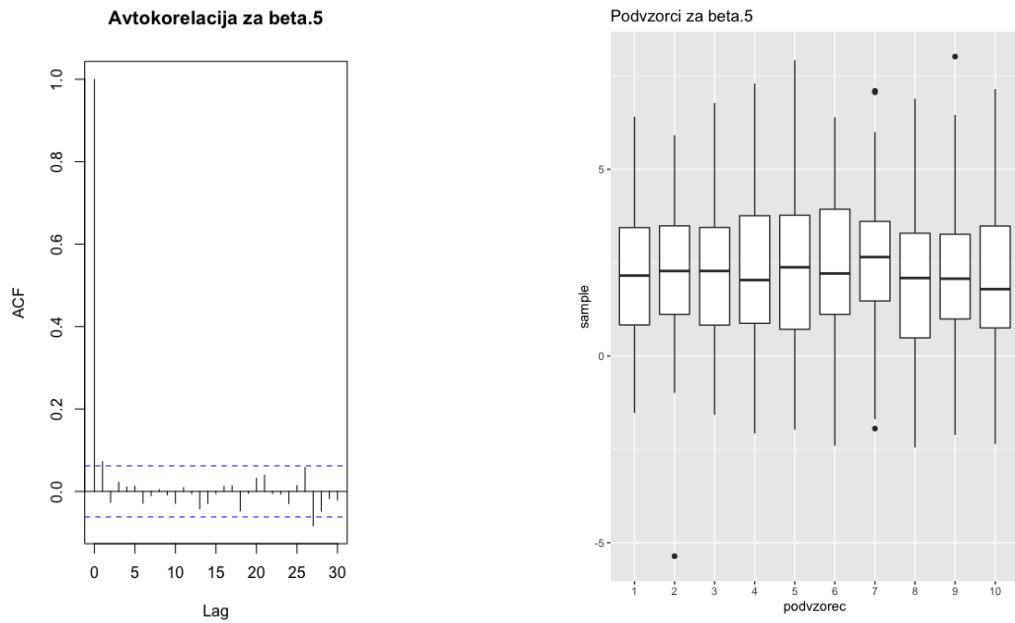


Figure 20: Grafa s prikazom avtokorelacijs in podvzorcev za koeficient β_5 .

Model z asimetrično porazdeljeno napako

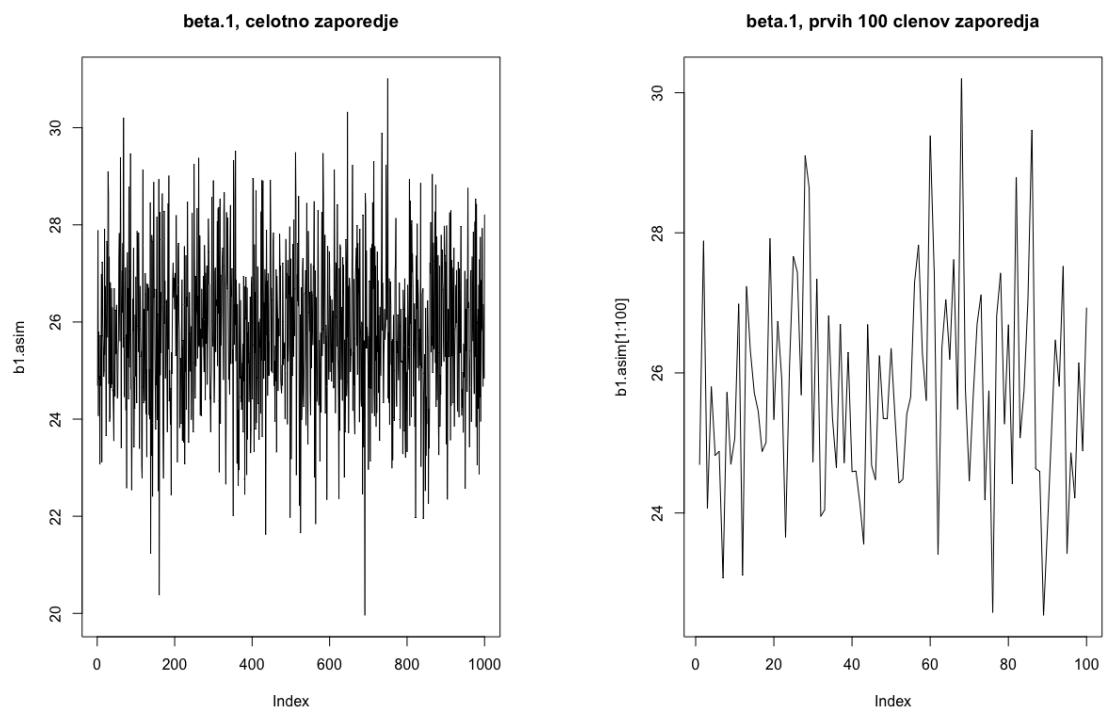


Figure 21: Grafa s prikazom konvergencije za koeficient β_1 .

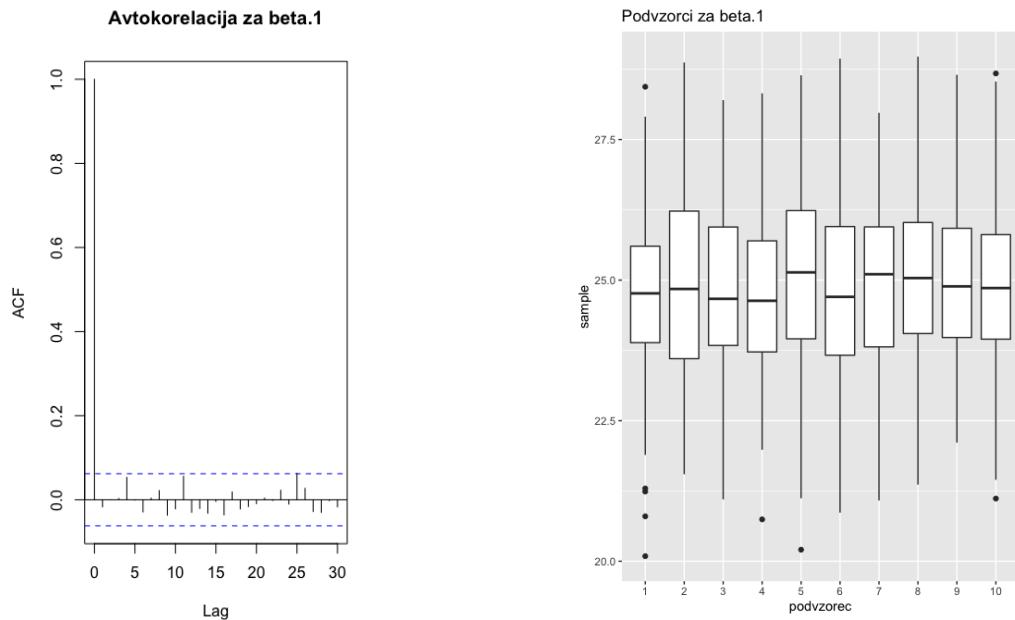


Figure 22: Grafa s prikazom avtokorelacijs in podvzorcev za koeficient β_1 .

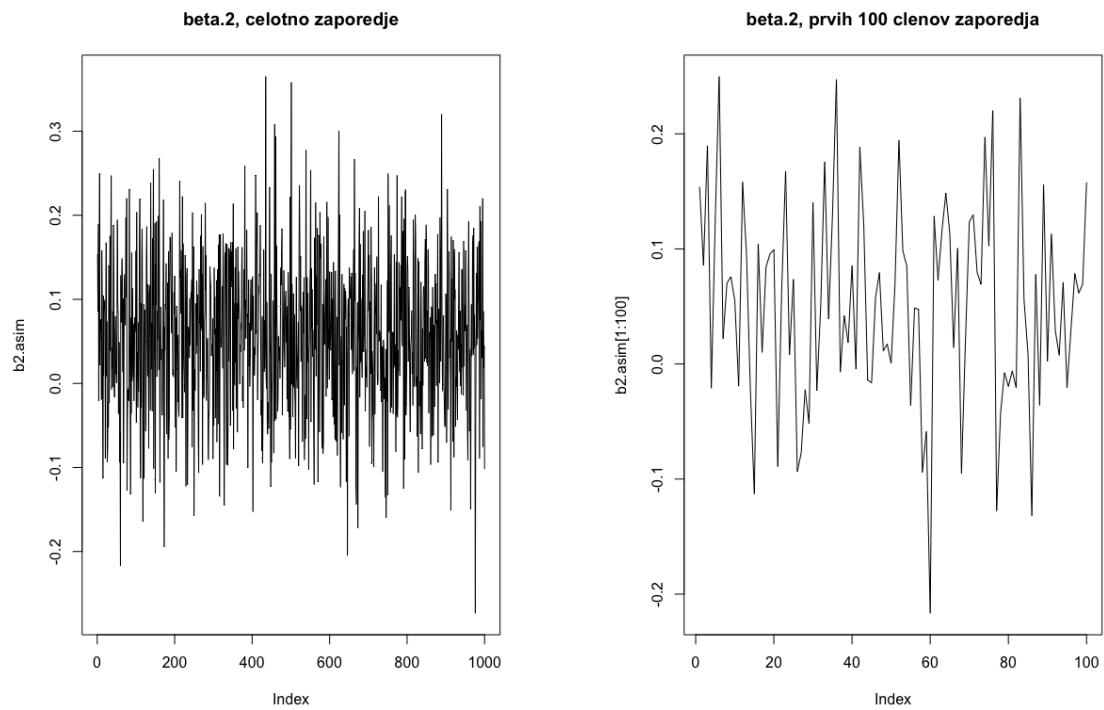


Figure 23: Grafa s prikazom konvergencije za koeficient β_2 .

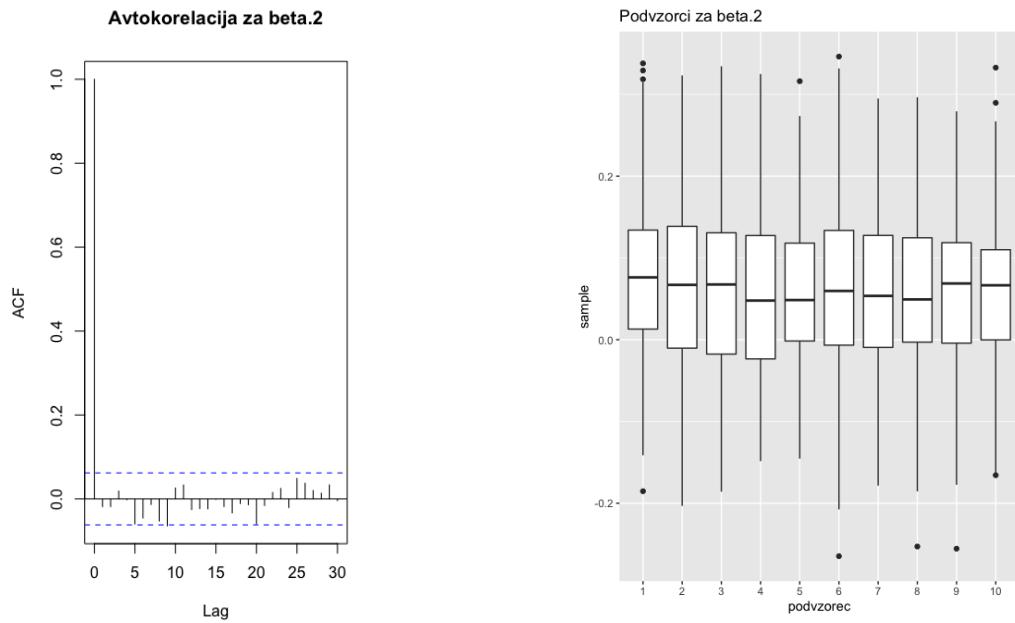


Figure 24: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_2 .

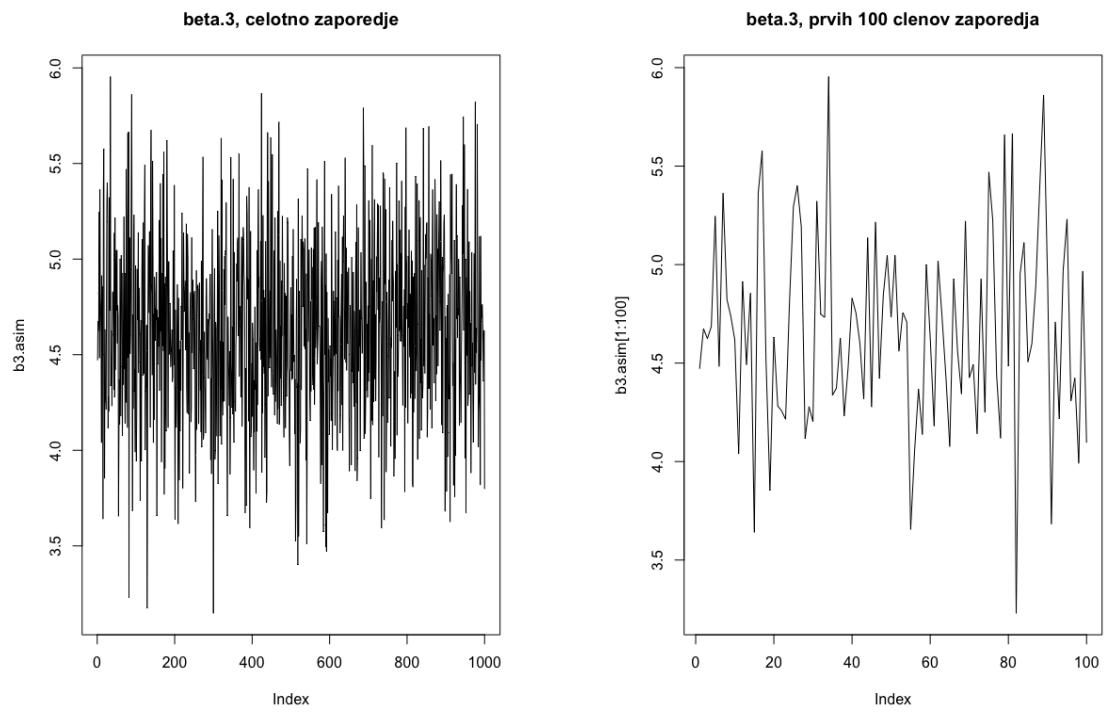


Figure 25: Grafa s prikazom konvergencije za koeficient β_3 .

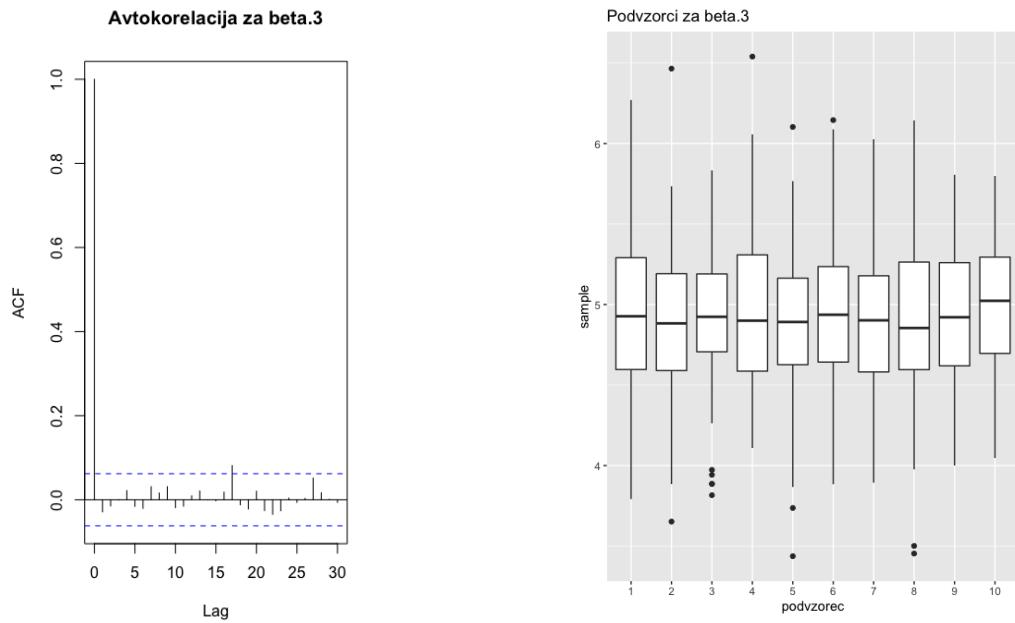


Figure 26: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_3 .

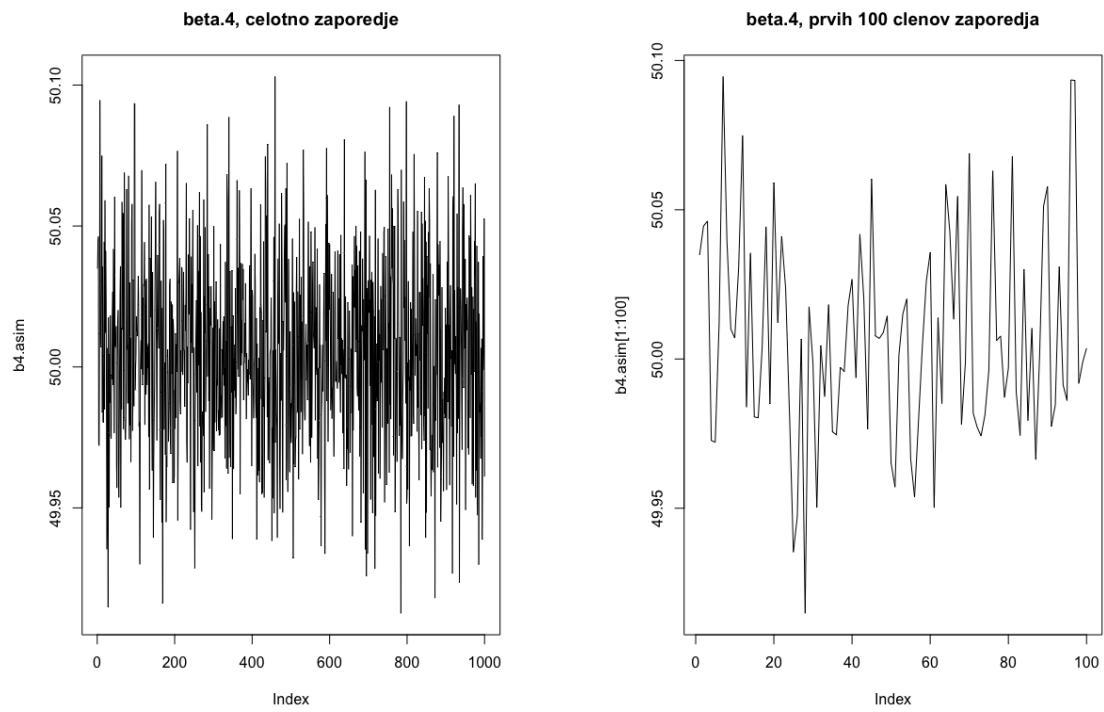


Figure 27: Grafa s prikazom konvergencije za koeficient β_4 .

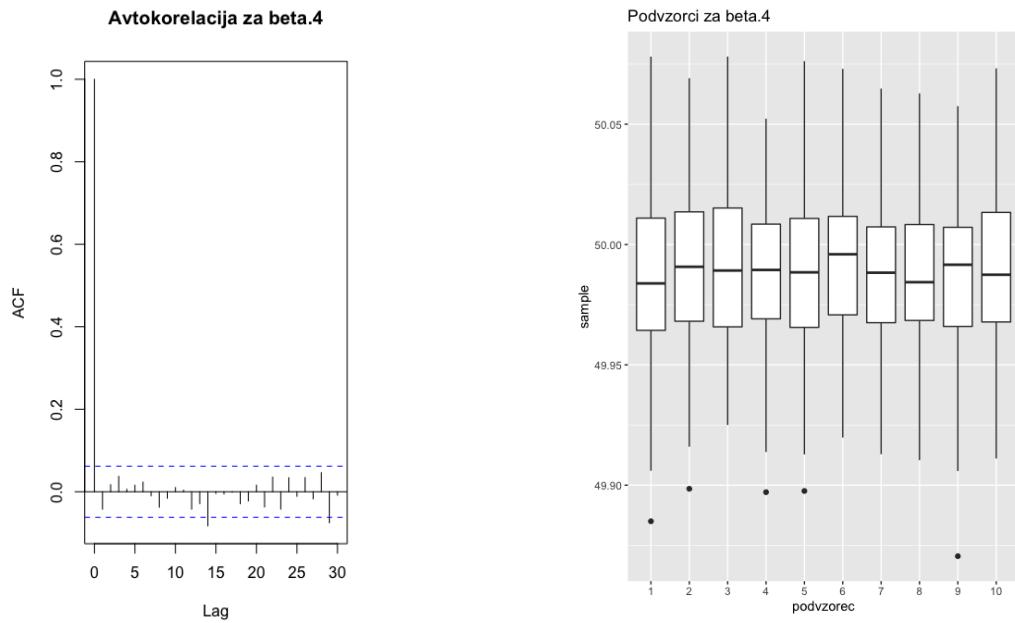


Figure 28: Grafa s prikazom avtokorelacijs in podvzorcev za koeficient β_4 .

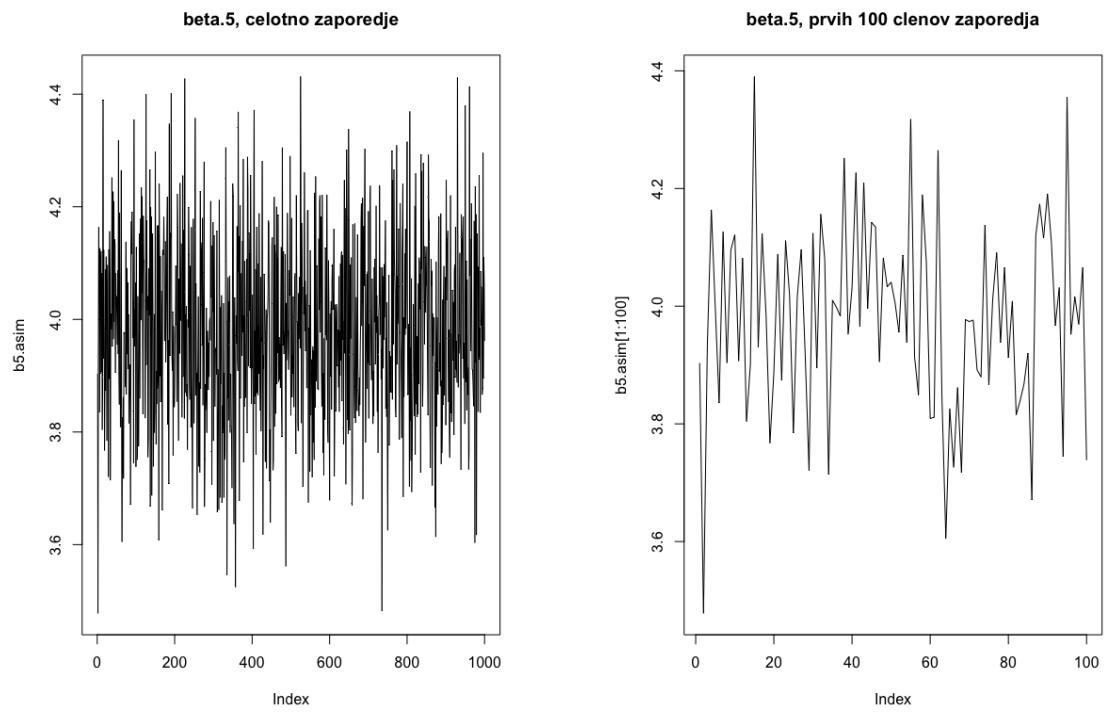


Figure 29: Grafa s prikazom konvergencije za koeficient β_5 .

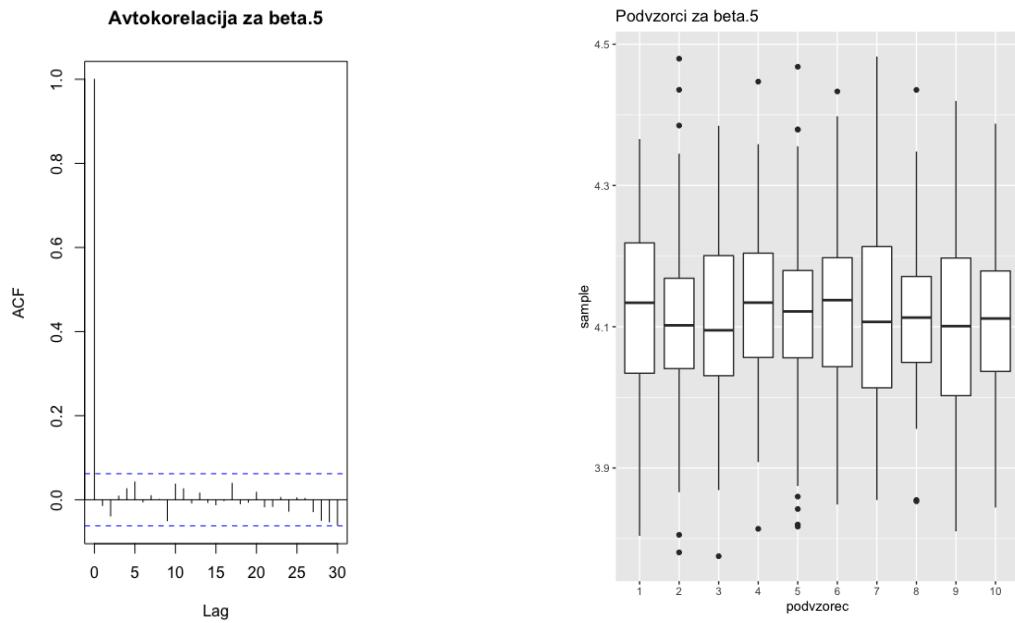


Figure 30: Grafa s prikazom avtokorelaciij in podvzorcev za koeficient β_5 .

Opazimo, da pri vseh treh modelih lahko rečemo, da pri izgledu verig ni težav. Prav tako so v redu tudi avtokorelaciije in podvzorci. S privzetimi nastavitevami funkcije `bayesx()` torej ni težav in jih zato ne spremojamo.

Primerjava Bayesovega pristopa s frekventističnim

Pri vsakem simulacijskem scenariju si za vsako simulacijo shranimo:

- Pri frekventističnem pristopu:
 - ocene regresijskih koeficientov,
 - ocene standardnih napak za regresijske koeficiente.
- Pri Bayesovem pristopu:
 - povrečne vrednosti za regresijske koeficient,
 - standardne odklone aposteriornih porazdelitev za regresijske koeficiente.

Za vsak simulacijski scenarij za oba pristopa poročamo:

- pristranskost,
- standardno napako in koren srednje kvadratne napake ocen,
- povprečje ocenjenih standardnih napak.

Opomba: V tabelah je prišlo do tipkarske napake, in sicer *koren standardne napake* namesto *koren srednje kvadratne napake*, ki pa je zaradi časovno potratnega poganjanja kode, nisem odpravila (morala bi vse simulacije še enkrat pognati, kar je trajalo veliko časa – za zajem tabele pa bi potem slikala del zaslona z izpisom).

Model z majhno varianco

Beta	Tocne.vrednosti.beta	Ocena.frevenstisticni	Pristanskost.frekventist	Ocena.Bayesov	Pristanskost.Bayesov
Intercept	25.000000	24.804112	0.195888	24.802744	0.197256
Ena	0.000000	-0.003222	0.003222	-0.003113	0.003113
Dva	5.000000	4.977617	0.022383	4.977387	0.022613
Tri	50.000000	50.001402	-0.001402	50.001445	-0.001445
Stiri	4.000000	3.997647	0.002353	3.997635	0.002365
Pet	10.000000	9.990063	0.009937	9.988621	0.011379

Figure 31: Tabela s prikazom pristranskosti za oba pristopa.

Beta	Tocne.vrednosti.beta	Standardna.napaka.frekventisticni	Koren.standardne.napake.frekventistični	Standardna.napaka.Bayesov	Koren.standardne.napake.Bayesov
Intercept	25.000000	1.721090	1.732202	1.725197	1.736438
Ena	0.000000	0.112635	0.112681	0.112483	0.112526
Dva	5.000000	0.552817	0.553270	0.554257	0.554719
Tri	50.000000	0.044544	0.044566	0.044536	0.044568
Stiri	4.000000	0.151895	0.151914	0.152202	0.152220
Pet	10.000000	0.537763	0.537855	0.538285	0.538405

Figure 32: Tabela s prikazom standardne napake in korena srednje napake za oba pristopa.

Beta	Tocne.vrednosti.beta	Povprečje.standardne.napake.frekventistični	Povprečje.standardne.napake.Bayesov
Intercept	25.000000	1.740489	1.994369
Ena	0.000000	0.110584	0.126635
Dva	5.000000	0.551796	0.632731
Tri	50.000000	0.044342	0.050816
Stiri	4.000000	0.151891	0.174344
Pet	10.000000	0.551038	0.631469

Figure 33: Tabela s prikazom povprečja ocenjenih standardnih napak pri obeh pristopih.

Opazimo, da ni večjih odstopanj med obema pristopoma. Največjo pristranskoost opazimo pri *interceptu*. Omenimo še, da je pristranskoost pri istem koeficientu z enakim predznakom v obeh pristopih.

Pri modelu z majhno varianco (malenkost) bolje simulira frekventistični model.

Model z veliko varianco

Beta	Tocne.vrednosti.beta	Ocena.frekvenstisticni	Pristranskoost.frekventist	Ocena.Bayesov	Pristranskoost.Bayesov
Intercept	25.000000	22.371669	2.628331	22.419489	2.580511
Ena	0.000000	-0.014458	0.014458	-0.022393	0.022393
Dva	5.000000	6.017601	-1.017601	6.000889	-1.000889
Tri	50.000000	49.885715	0.114285	49.891488	0.108512
Stiri	4.000000	4.007030	-0.007030	4.000626	-0.000626
Pet	10.000000	9.117465	0.882535	9.129321	0.870679

Figure 34: Tabela s prikazom pristranskosti za obo pristopa.

Beta	Tocne.vrednosti.beta	Standardna.napaka.frekventistični	Koren.standardne.napake.frekventistični	Standardna.napaka.Bayesov	Koren.standardne.napake.Bayesov
Intercept	25.000000	85.813918	85.854159	85.756729	85.795545
Ena	0.000000	5.495247	5.492942	5.492988	5.492988
Dva	5.000000	27.934005	27.952534	27.917933	27.935869
Tri	50.000000	2.187168	2.190152	2.187396	2.190886
Stiri	4.000000	7.632750	7.632753	7.632369	7.632369
Pet	10.000000	28.012630	28.026529	28.032142	28.045661

Figure 35: Tabela s prikazom standardne napake in korena srednje napake za obo pristopa.

Beta	Tocne.vrednosti.beta	Povprečje.standardne.napake.frekventistični	Povprečje.standardne.napake.Bayesov
Intercept	25.000000	88.198007	89.209388
Ena	0.000000	5.549838	5.613451
Dva	5.000000	27.829217	28.113251
Tri	50.000000	2.231778	2.257243
Stiri	4.000000	7.711650	7.799341
Pet	10.000000	27.858309	28.188683

Figure 36: Tabela s prikazom povprečja ocenjenih standardnih napak pri obeh pristopih.

Podobno kot prej tudi tukaj najbolj odstopa od prave vrednosti koeficient *intercept*, vsi enaki koeficienti pa imajo ponovno enako predznačeno pristranskoost.

Tudi tukaj sta obo pristopa zelo izenačena, a je malo boljši frekventistični (ni pa opaziti nobenih večjih razlik).

Komentirajmo še, da so napake pri *interceptu* tukaj malenkost večje, verjetno prav zaradi ve-

likosti izbrane velike variance na začetku. Verjetno bi varianco lahko zmanjšali nekje na 50 ali še malenkost manj.

Model z asimetrično porazdeljeno napako

Beta	Tocne.vrednosti.beta	Ocena.frekvenstisticni	Pristranskost.frekventist	Ocena.Bayesov	Pristranskost.Bayesov
Intercept	25.000000	25.992251	-0.992251	25.992499	-0.992499
Ena	0.000000	0.000585	-0.000585	0.000578	-0.000578
Dva	5.000000	5.004164	-0.004164	5.004290	-0.004290
Tri	50.000000	50.001200	-0.001200	50.001165	-0.001165
Stiri	4.000000	4.005963	-0.005963	4.005989	-0.005989
Pet	10.000000	9.995056	0.004944	9.994828	0.005172

Figure 37: Tabela s prikazom pristranskosti za oba pristopa.

Beta	Tocne.vrednosti.beta	Standardna.napaka.frekventisticni	Koren.standardne.napake.frekventisti	Standardna.napaka.Bayesov	Koren.standardne.napake.Bayesov
Intercept	25.000000	1.101783	1.482730	1.101733	1.482859
Ena	0.000000	0.069556	0.069558	0.069491	0.069493
Dva	5.000000	0.353352	0.353376	0.354202	0.354228
Tri	50.000000	0.027648	0.027674	0.027653	0.027678
Stiri	4.000000	0.100125	0.100302	0.100157	0.100336
Pet	10.000000	0.344124	0.344159	0.344883	0.344922

Figure 38: Tabela s prikazom standardne napake in korena srednje napake za oba pristopa.

Beta	Tocne.vrednosti.beta	Povprečje.standardne.napake.frekventisticni	Povprečje.standardne.napake.Bayesov
Intercept	25.000000	1.098862	1.460335
Ena	0.000000	0.069584	0.092510
Dva	5.000000	0.346833	0.461166
Tri	50.000000	0.027682	0.036734
Stiri	4.000000	0.095596	0.127292
Pet	10.000000	0.346635	0.461385

Figure 39: Tabela s prikazom povprečja ocenjenih standardnih napak pri obeh pristopih.

Tudi tukaj veljajo podobne opazke kot pri prejšnjih dveh primerih. Opazimo še, da je pri koeficientu *intercept* absolutna vrednost pristranskosti enaka 1 (sicer v minus), kar je približno enako kot je povprečje standardne napake.

Iz zgornje analize je sicer težko povzeti kakšne globoke zaključke. Verjetno bi lahko kaj več videli in komentirali, če bi pognali večje število simulacij.

Na podlagi samo teh podatkov bi rekla, da sta oba pristopa približno enako dobra, če gledamo samo na rezultate, bi pa v primeru linearne regresije izbrala frekventistični pristop, saj so stvari lažje izračunljive in manj potratne. Bayesov pristop je po mojem mnenju boljši za reševanje kakšnih drugačnih vrst problemov. Sama bi torej vedno, ko sta pristopa primerljivo dobra, izbrala frekventistični pristop.