

Magistrski študij Uporabna statistika  
**Bayesova statistika**  
**2020/2021**  
**Seminarska naloga**

Seminarsko nalogo rešujte samostojno.

Izdelajte R Markdown poročilo, v katerem naj bo celotna koda, izpisi rezultatov in komentarji (poročilo naj bo zaključena celota).

- Datoteka Rmd naj vsebuje takšno kodo, da bo delovalo branje podatkov iz datotek, ki so v isti mapi kakor Rmd datoteka.
- Celotna koda mora delovati na operacijskem sistemu Windows.
- Če je za kakšno simulacijo oz. vzorčenje potrebnega veliko časa, potem si lahko rezultate shranite v posebno R datoteko (funkcija `dump`) in jih nato v poročilo uvozite (funkcija `source`), pri tem pa v poročilo vključite kodo za izračun ter v tistem delu uporabite opcijo `eval = FALSE`.
- Oddajte izvirno Rmd datoteko (poimenujte jo v smislu SN\_priimek), njen izpis v pdf ali html formatu (v slednjem primeru ne uporabljajte šumnikov), datoteke podatkov in morebitne datoteke rezultatov simulacij.

1. **Pri multipli linearni regresiji preko simulacij primerjajte Bayesov pristop s frekventističnim.** Pri obeh pristopih uporabite običajni normalni model: pri Bayesovem pristopu uporabite funkcijo `bayesx`, pri frekventističnem pa `lm`. Primerjavo pristopov naredite glede na tri različne scenarije: normalno porazdeljena napaka z majhno varianco, normalno porazdeljena napaka z veliko varianco, asimetrično porazdeljena napaka (npr. hi-kvadrat). Pri slednjem scenariju boste primerjali Bayesov in frekventistični pristop takrat, ko predpostavke modelov ne držijo – preverjali boste njuno robustnost. Pri vsakem izmed treh scenarijev naj bodo preostali parametri simulacij enaki, za njih pa naj velja:
  - Število ponovitev simulacij naj bo med 1000 in 10.000 (raje več kot manj, odvisno pa od zmoglosti vašega računalnika).
  - Velikost vzorca naj bo 100.
  - Pojasnjevalne spremenljivke simulirajte neodvisno eno od druge. Simulirate jih lahko iz različnih porazdelitev (normalna, Bernoullijeva ipd.). Za spremenljivke si izberite različno močne efekte: vsaj ena spremenljivka naj ima močan efekt, vsaj ena šibek efekt, vsaj ena pa naj bo brez efekta (regresijski koeficient enak nič). Kakšen efekt je močan in kakšen šibek, je seveda odvisno od vrednosti spremenljivk (preizkusite nekaj različnih možnosti, v poročilo pa vključite le eno kombinacijo regresijskih koeficientov).
  - Poleg petih neodvisnih spremenljivk uporabite v linearnem modelu tudi neko začetno vrednost (*intercept*).
  - Pri nekaj simulacijah preverite, ali so privzeti parametri funkcije `bayesx` primerni (število iteracij, *burn-in*, *thinning*), da je konvergenca zadovoljiva. Po potrebi jih spremenite, znotraj enega simulacijskega scenarija pa uporabljajte enake nastavitve. Kratko poročajte o tem.

Sami si izberite manjšo in večjo varianco pri normalni porazdelitvi napak (kakšni bosta, bo zopet odvisno od vrednosti spremenljivk). Prav tako si sami izberite vrsto asimetrične porazdelitve napak pri tretjem simulacijskem scenariju.

Pri vsakem simulacijskem scenariju si ob vsaki simulaciji zapomnite naslednje:

- Pri frekventističnem pristopu si zapomnite ocene regresijskih koeficientov in ocene njihovih standardnih napak.
- Pri Bayesovem pristopu si zapomnite povprečne vrednosti in standardne odklone aposteriornih porazdelitev za regresijske koeficiente, tj. za oceno posameznega regresijskega koeficienta bomo vzeli povprečje aposteriorne porazdelitve, medtem ko bo standardni odklon aposteriorne porazdelitve meril standardno napako cenilke.

Za vsak simulacijski scenarij poročajte in med seboj ustrezno primerjajte ter komentirajte naslednje:

- pristranost (*bias*) pri obeh pristopih,
- standardno napako in koren srednje kvadratne napake ocen pri obeh pristopih,
- povprečje ocenjenih standardnih napak pri obeh pristopih.

2. **Izberite si neke podatke.** Vsebujejo naj vsaj pet spremenljivk in vsaj 50 statističnih enot. Podatki naj bodo taki, da jih vsaj nekoliko razumete, tako da boste lahko postavili kakšno smiselno raziskovalno vprašanje in jih interpretirali.

- (a) Na kratko predstavite podatke in zapišite, kje ste jih dobili. Poročajte o morebitnem čiščenju podatkov.
- (b) Postavite linearen model z vsaj dvema neodvisnima spremenljivkama.
  - Na kratko razložite, zakaj ste si ga izbrali.
  - Ocenite model z metodami Bayesove statistike. Preucite konvergenco.
  - Interpretirajte dobljene rezultate (ocena, *credible interval*, statistična značilnost). Interpretacija naj bo vsebinsko polna.
- (c) Ocenite nek hierarhični model, preucite konvergenco, interpretirajte rezultate.