# Osnutek seminarske naloge RZM

Anja Žavbi Kunaver in Vesna Zupanc

10 12 2020

## Ideja naloge

Obravnavana metoda je linearna regresija in zanima nas, kako kršenje predpostavk (konkretneje nenormalna porazdelitev ostankov in močna korelacija med pojasnjevalnimi spremenljivkami) vpliva na njene rezultate. Preverjali bomo vpliv različnih dejavnikov na pristranost in širino intervalov zaupanja regresijskih koeficientov. Za izračun intervalov zaupanja bomo uporabili različne metode, ki smo jih spoznali v poglavjih simulacij in metod samovzorčenja. Tako bomo primerjali pokritosti in širine različno ocenjenih intervalov zaupanja. Poleg tega bomo preverjali, če probleme kršenja predpostavk lahko (vsaj delno) odpravimo z uporabo posplošenih modelov.

## Generiranje podatkov

Fiksni parametri pri generiranju podatkov so sledeči:

- porazdelitev pojasnjevalnih spremenljivk:
  - $X_1 \sim Gamma(2, 5)$
  - $X_2 \sim Gamma(2, 5)$
  - $X_3 \sim Gamma(5, 5)$
  - $X_4 \sim Gamma(5, 5)$
  - $X_5 \sim Gamma(5, 5)$
- formula za generiranje podatkov:

$$y_i = 5x_1 + x_2 + 5x_3 + x_4 + 0x_5 + \epsilon_i$$

Pri generiranju podatkov bomo spreminjali sledeče:

- velikost vzorca $n \in \{10, 20, 30, 50, 100, 500, 1000\}$
- korelacija med odvisnimi spremenljivkami ($cor \in \{0, 0.3, 0.6, 0.9\}$)
- porazdelitev napak ($Gamma(\alpha, \beta)$), kjer bomo parameter $\alpha$ spreminjali tako, da dobimo različno močno asimetrične porazdelitve ($(\alpha, \beta) \in \{(1, 5), (2, 5), (2, 2), (5, 5)\}$
- v modelu ne upoštevamo vseh neodvisnih spremenljivk (spreminjamo število spremenljivk, ki jih upoštevamo)

Pri generiranju koreliranih gama spremenljivk uporabimo sledečo lastnost: Če $X_i \sim Gamma(k_i, \theta)$, potem je

$$\sum_{i=1}^{n} X_i \sim Gamma(\sum_{i=1}^{n} k_i, \theta)$$

## Potek dela

Za vsako kombinacijo podatkov bomo naredili po 1000 simulacij, kjer bomo na vsakem koraku simulacije generirali podatke in nato ocenili model s pomočjo linearne regresije. Na vsakem koraku bomo shranili oceno regresijskih koeficientov in na koncu ocenili intervale zaupanja (enkrat naivni, drugič na podlagi standardnih napak, tretjič obrnjeni).

Iz intervalov zaupanja bomo dobili pristranost ocen regresijskih koeficientov ter seveda širine intervalov zaupanja. Podatke bomo grafično prikazali v odvisnosti od parametrov simulacij in s pomočjo tega poskušali ugotoviti, kako kršenje predpostavk vpliva na linearno regresijo ter kakšen vpliv ima pri različno velikih vzorcih in različnem številu vključenih spremenljivkah.

Na istih podatkih bomo postopek ponovili z uporabo posplošenih linearnih modelov (funkcija `glm()`), za katere vemo, da se bolje obnesejo ob nenormalno porazdeljenih odvisnih spremenljivkah, ter preverili, ali s tem lahko odpravimo (morebitne) težave zaradi kršenja predpostavk.

Za predstavitev in primerjavo rezultatov bomo uporabili grafične prikaze in tabele ter vse skupaj poskušali podkrepiti z literaturo.