Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

11.01.2020

Uvod

- Opazujemo širino in pokritost IZ regresijskih koeficientov
- Ocenjevanje intervalov zaupanja s funkcijo `confint.default()`
- Linearna regresija (lm) in posplošeni linearni modeli (glm)
- Metoda najmanjših kvadratov (OLS) in metoda iterativnega uteženega povprečja najmanjših kvadratov (IWLS)
- Prednost funkcije `glm()`: `family, link`

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

Predpostavke LR

- linearnost regresijskega modela: $y = \beta_1 + \beta_2 x_i + u_i$
- ničelna povprečna vrednost u_i: $E(u_i) = 0$
- homoskedastičnost: $Var(u_i) = E(u_i^2) = \sigma^2$
- odsotnost avtokorelacije: $cov(e_i, e_j | x_i, x_j) = 0$ za vsak $i \neq j$
- $Cov(x_2, u) = Cov(x_3, u) = ... = Cov(x_k, u) = 0$
- število opazovanj mora presegati število ocenjenih parametrov
- $Var(X)$ je končno pozitivno število
- pravilno specificiran regresijski model
- odsotnost multikolinearnosti: $\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0$
- $u_i \sim N(0, \sigma_u^2)$.

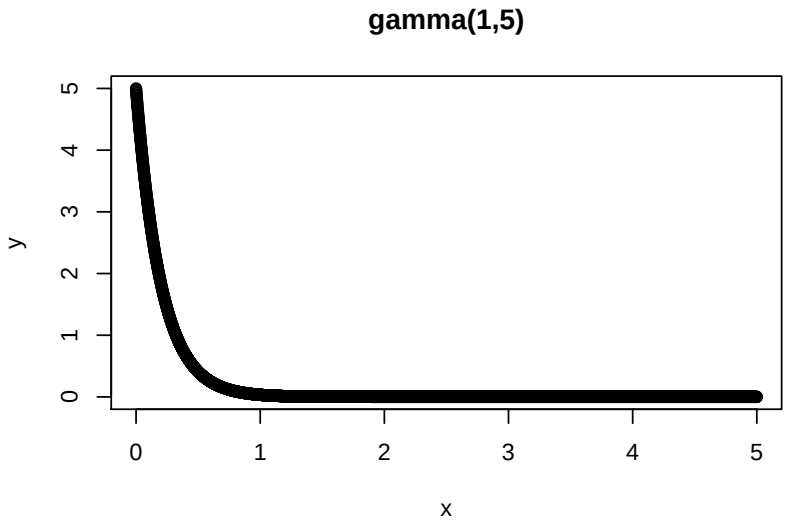
Generiranje podatkov

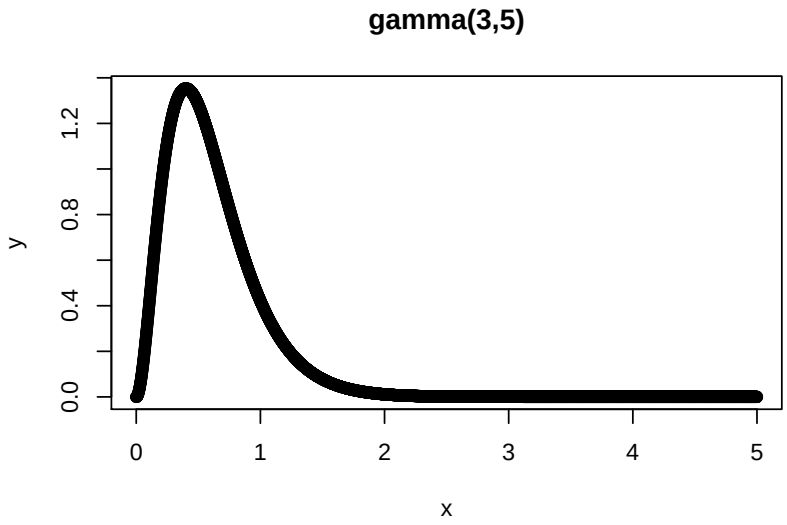
- formula za generiranje podatkov:

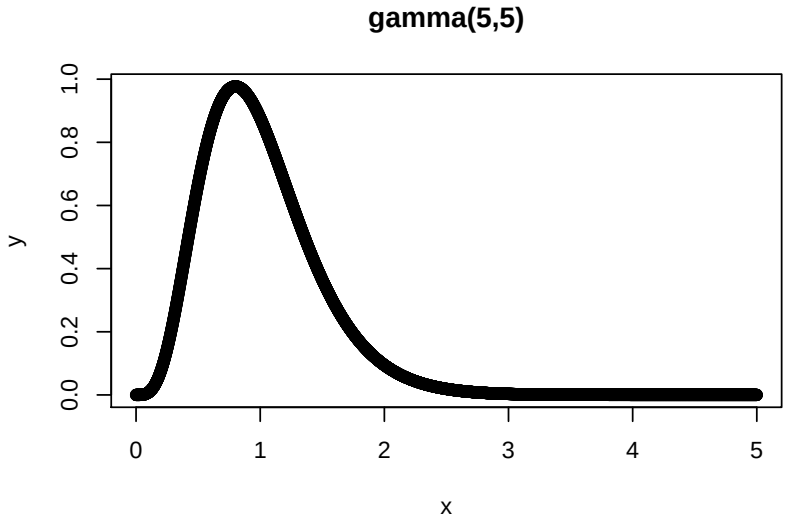
$$y_i = 1 + x_1 + x_2 + 0x_3 + \epsilon_i.$$

- velikost vzorca $n \in \{10, 50, 100, 500, 1000\}$;
- korelacija med pojasnjevalnimi spremenljivkami ($cor \in \{0, 0.3, 0.6, 0.9\}$);
- porazdelitev pojasnjevalnih spremenljivk: $X_j \sim Gamma(\delta, 5)$, $j = 1, 2, 3$, $\delta = 2, 5$;
- porazdelitev napak $Gamma(\alpha, 5)$, $\alpha \in \{1, 3, 5\}$;
- v modelu ne upoštevamo vseh neodvisnih spremenljivk: enkrat vključimo vse spremenljivke, enkrat izločimo X_3, enkrat pa X_2.

Porazdelitev gama

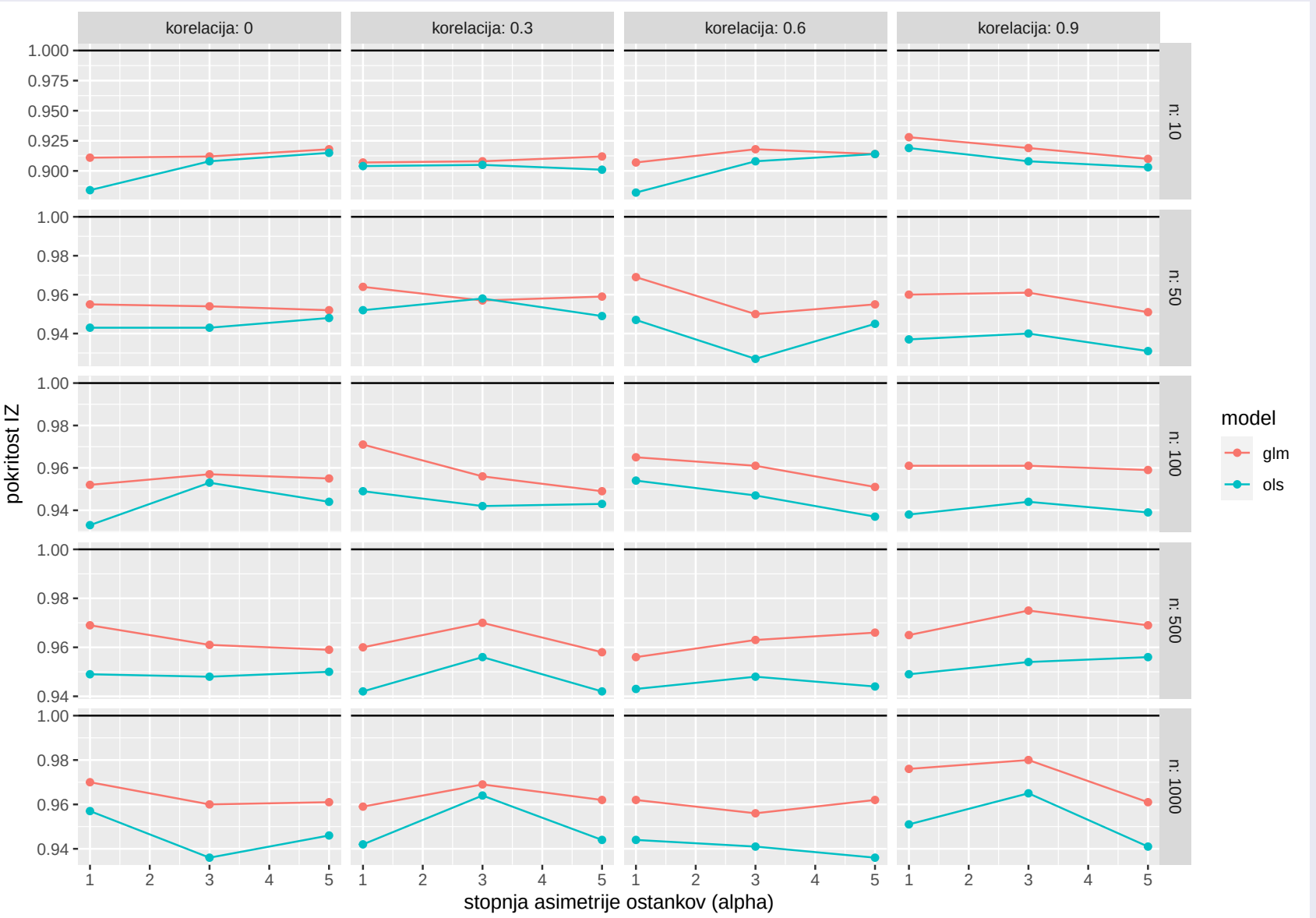






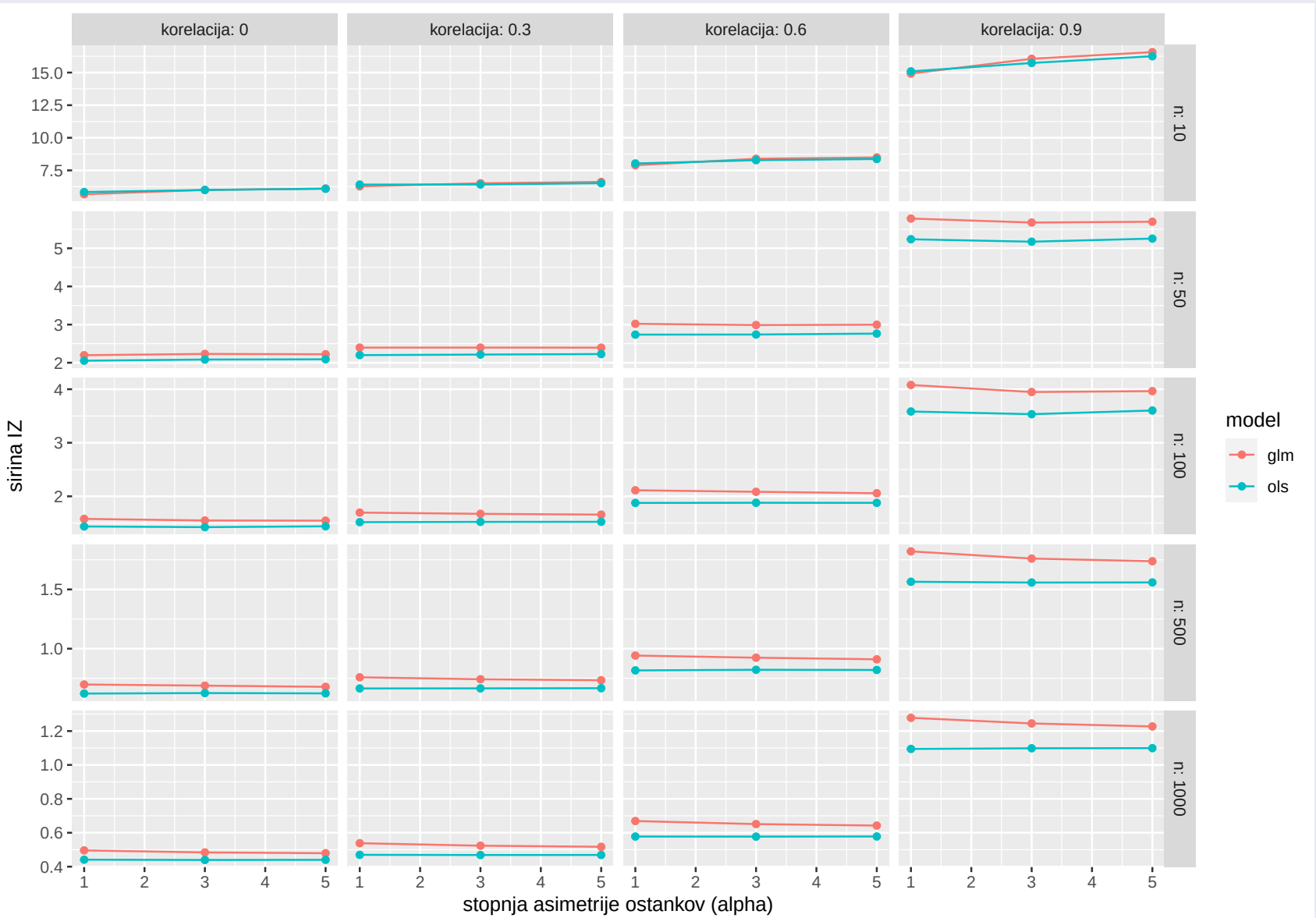
Pregled rezultatov - Polni model

X ~ Gamma(2,5) - pokritost



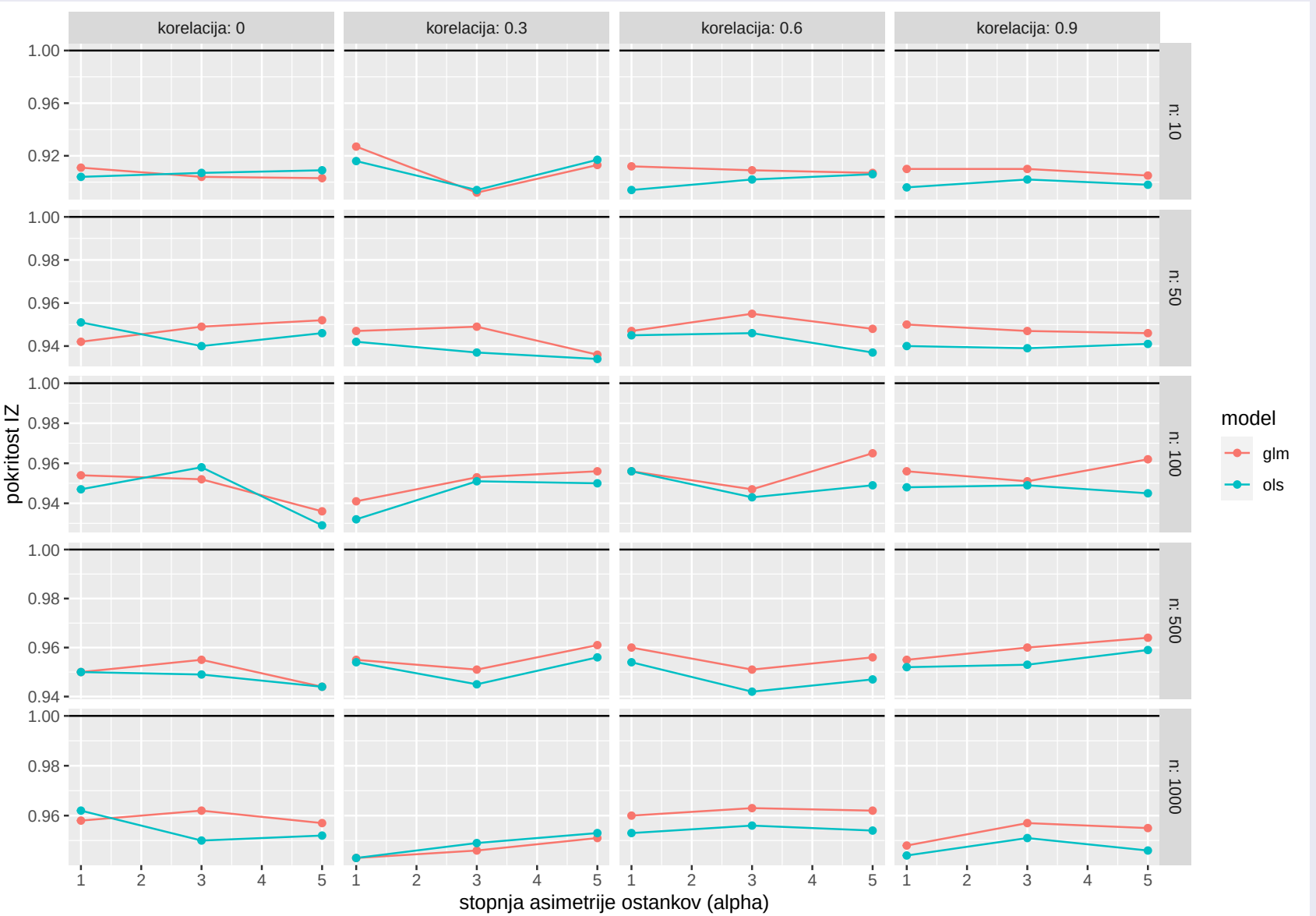
Pregled rezultatov - Polni model

X ~ Gamma(2,5) - širina IZ



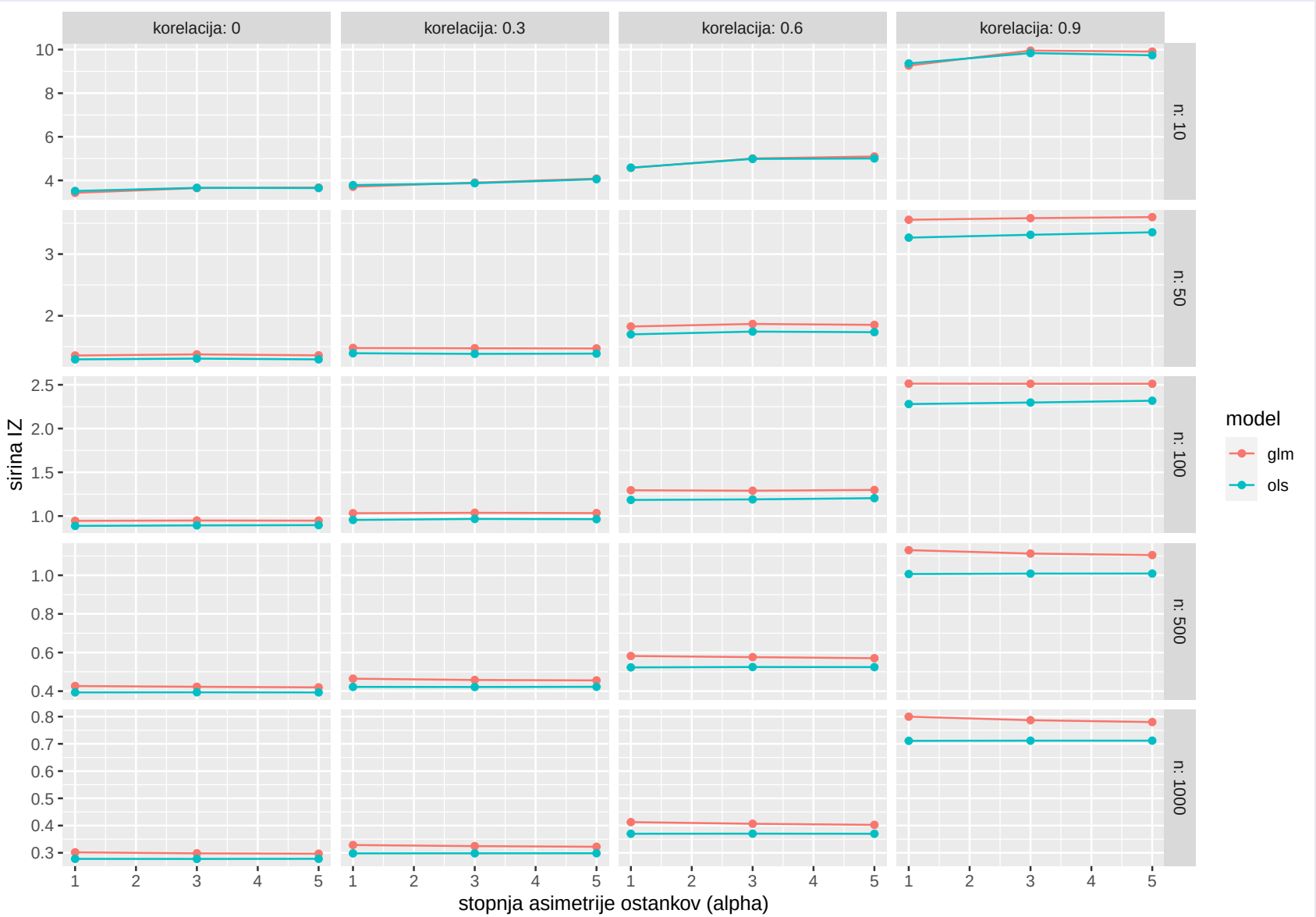
Pregled rezultatov - Polni model

X ~ Gamma(5,5) - pokritost



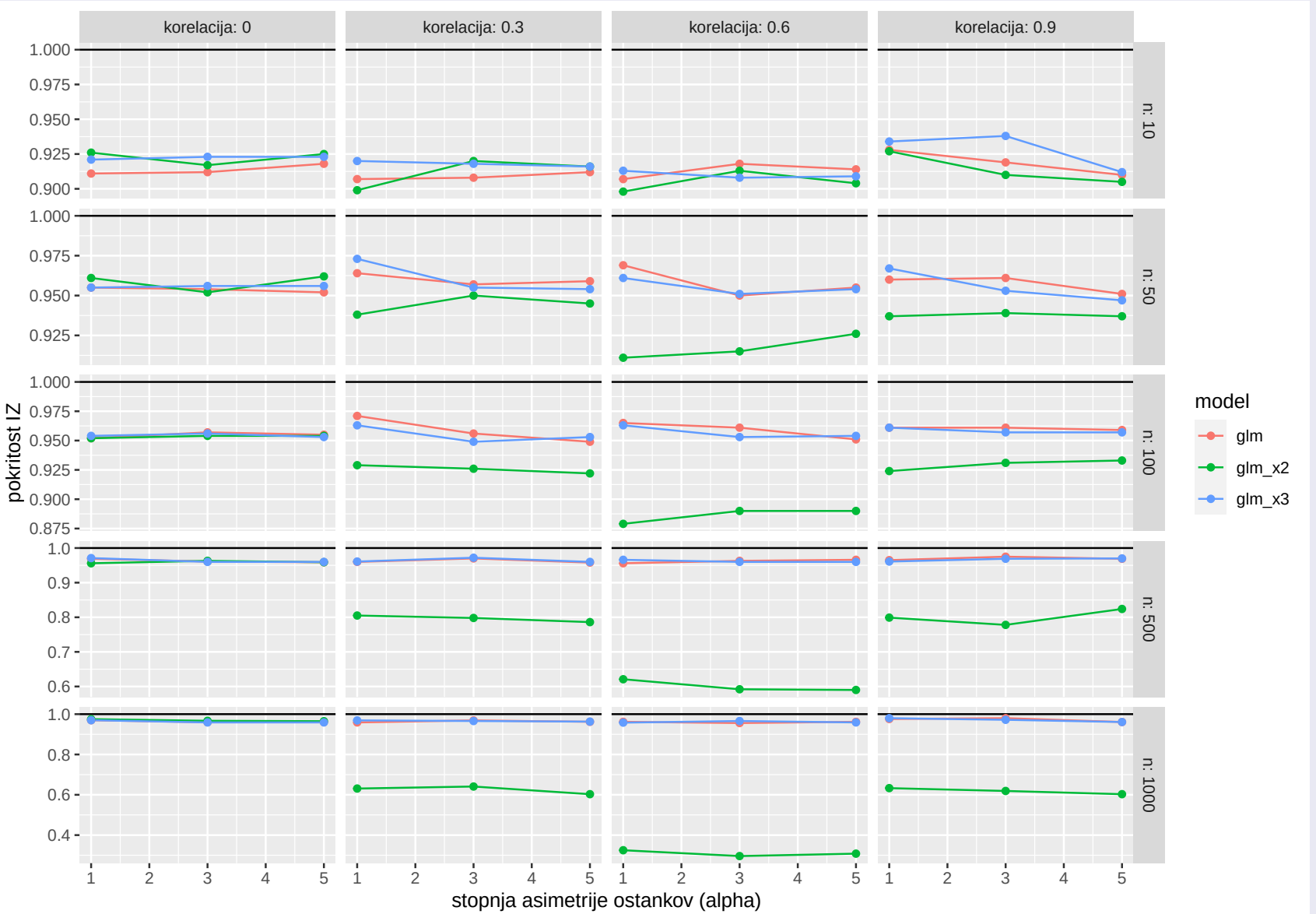
Pregled rezultatov - Polni model

X ~ Gamma(5,5) - širina IZ



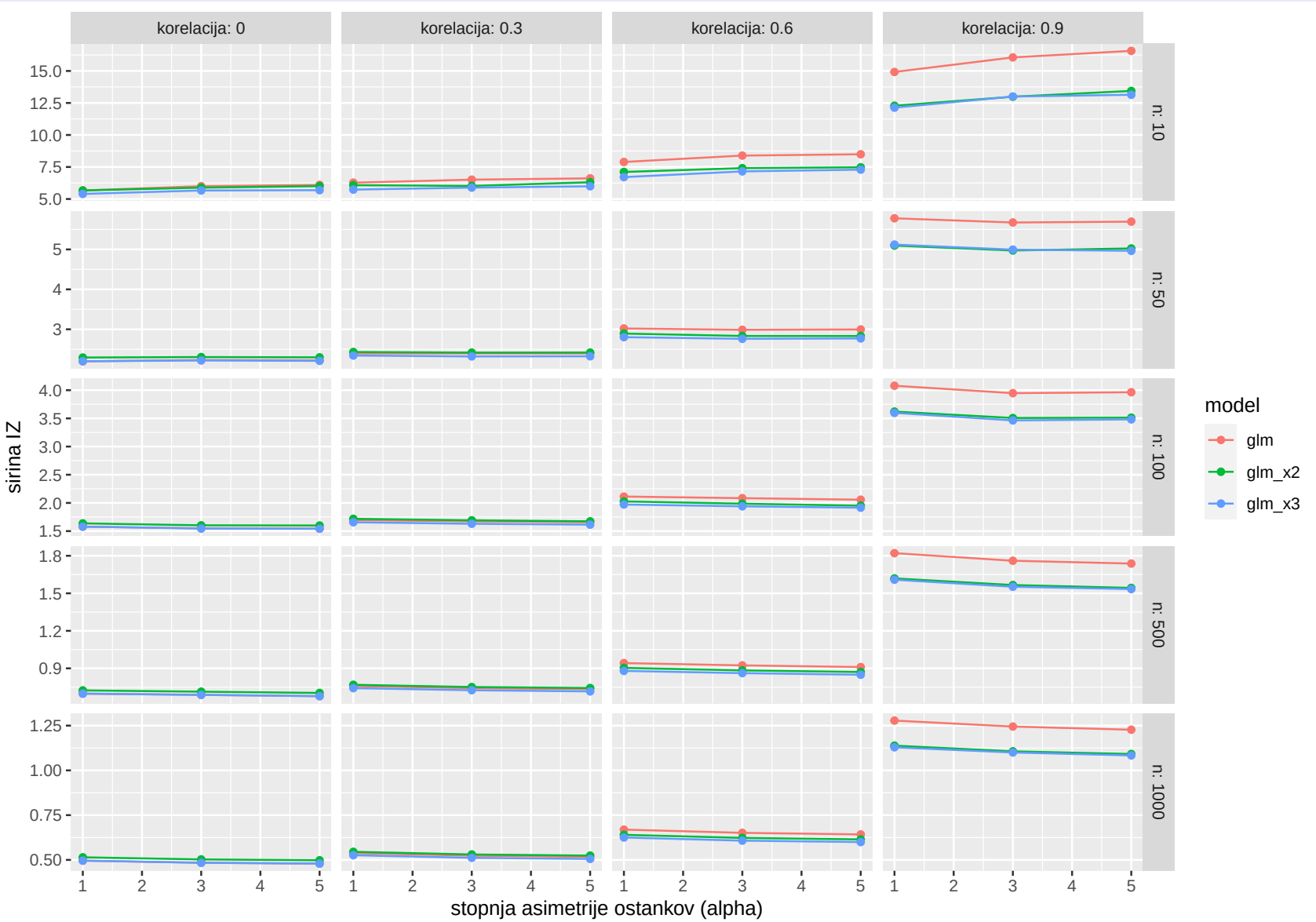
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(2,5) - GLM pokritost



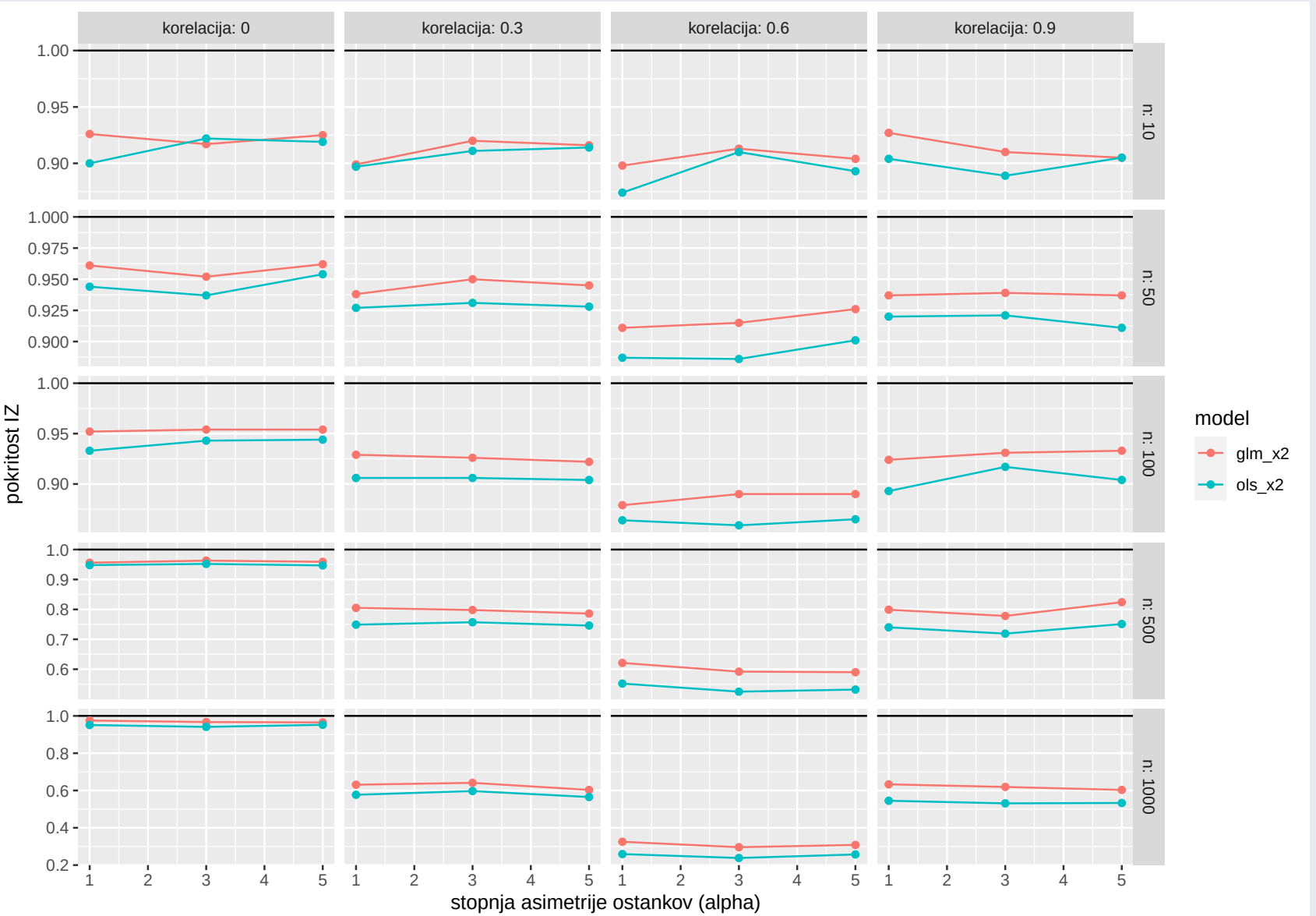
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(2,5) - GLM širina IZ



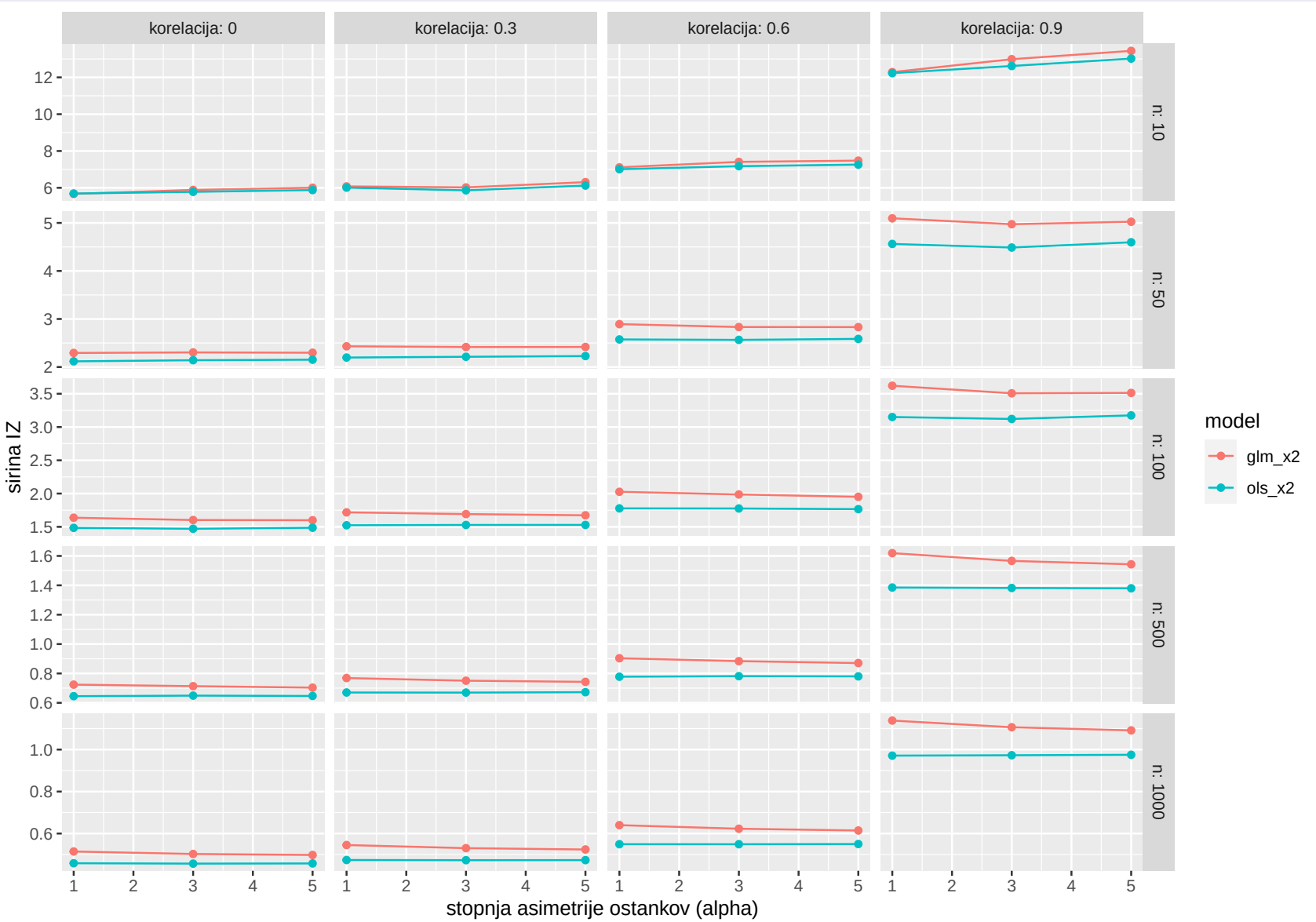
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(2,5) - primerjava GLM in OLS brez X2 - pokritost



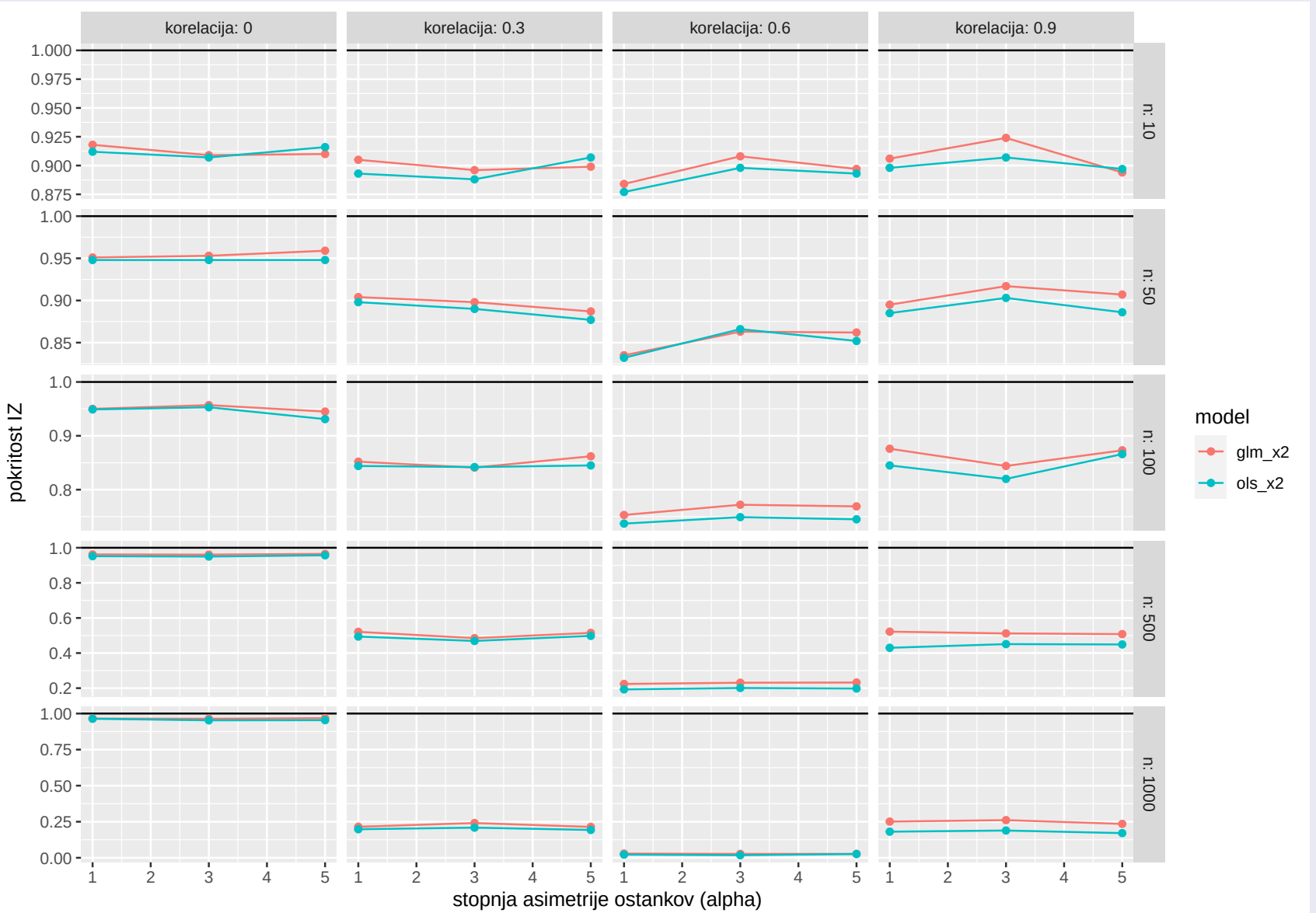
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(2,5) - primerjava GLM in OLS brez X2 - širina IZ



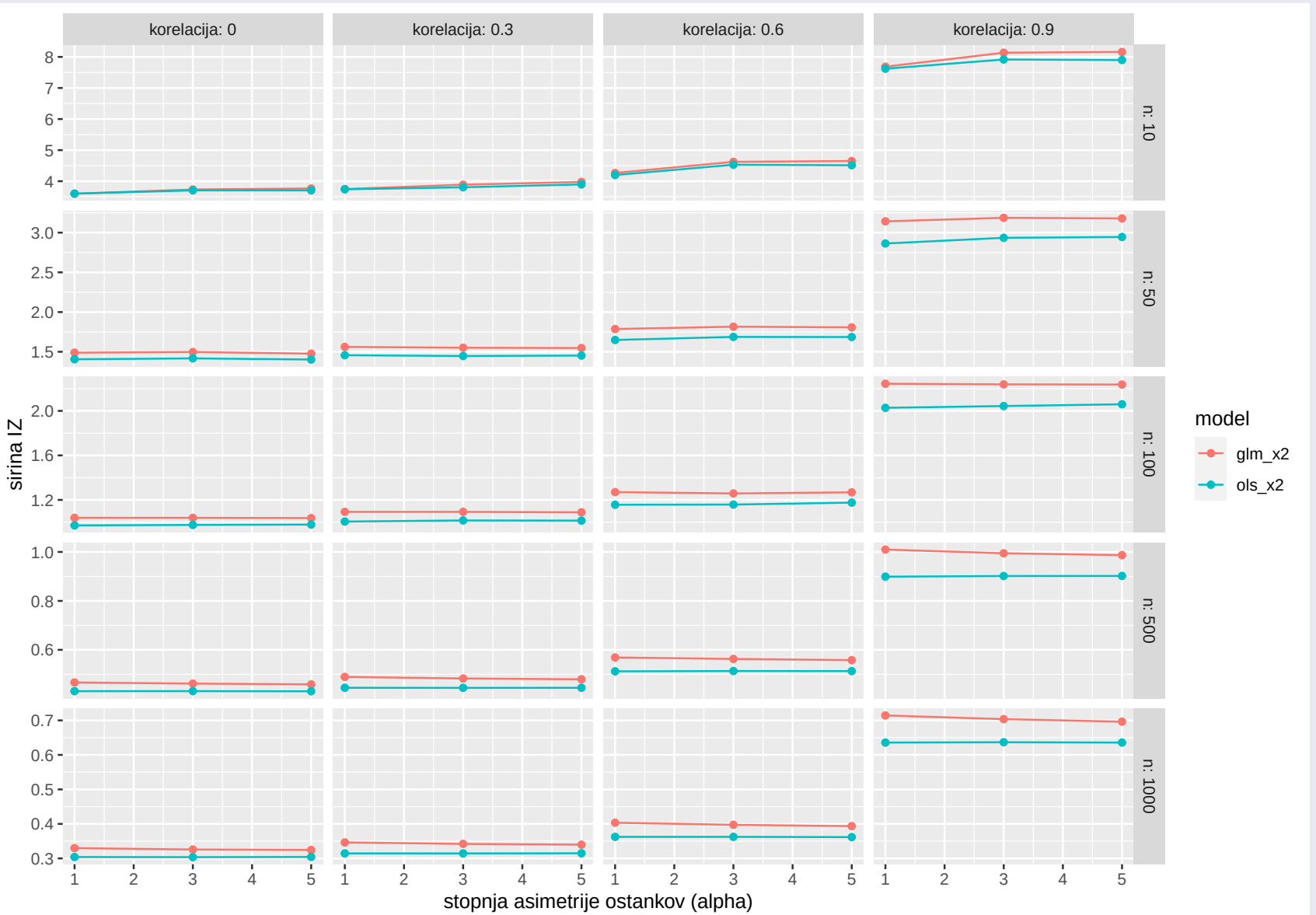
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(5,5) - primerjava GLM in OLS brez X2 - pokritost



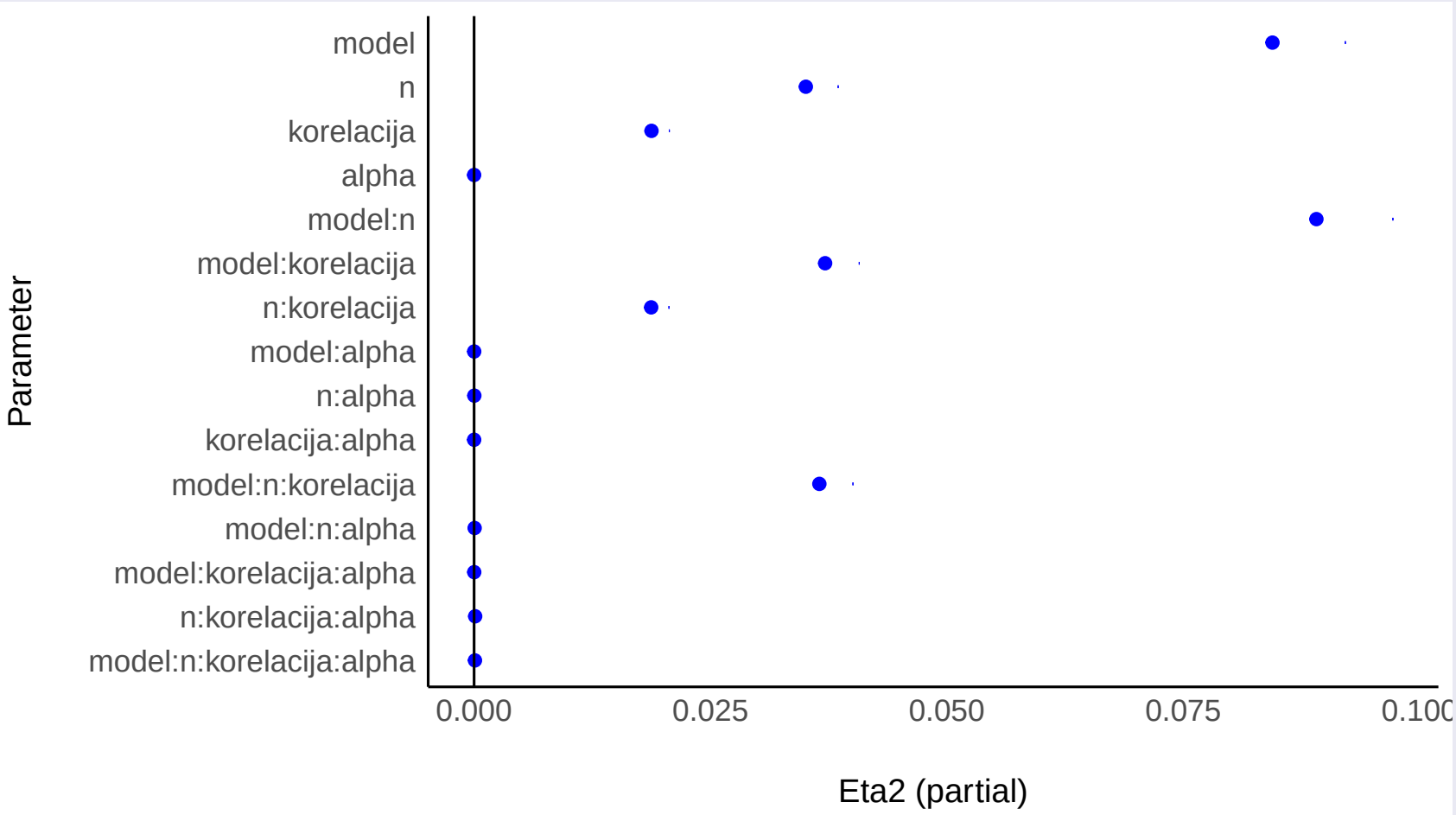
Pregled rezultatov - Odstranjevanje spremenljivk

X ~ Gamma(5,5) - primerjava GLM in OLS brez X2 - širina IZ



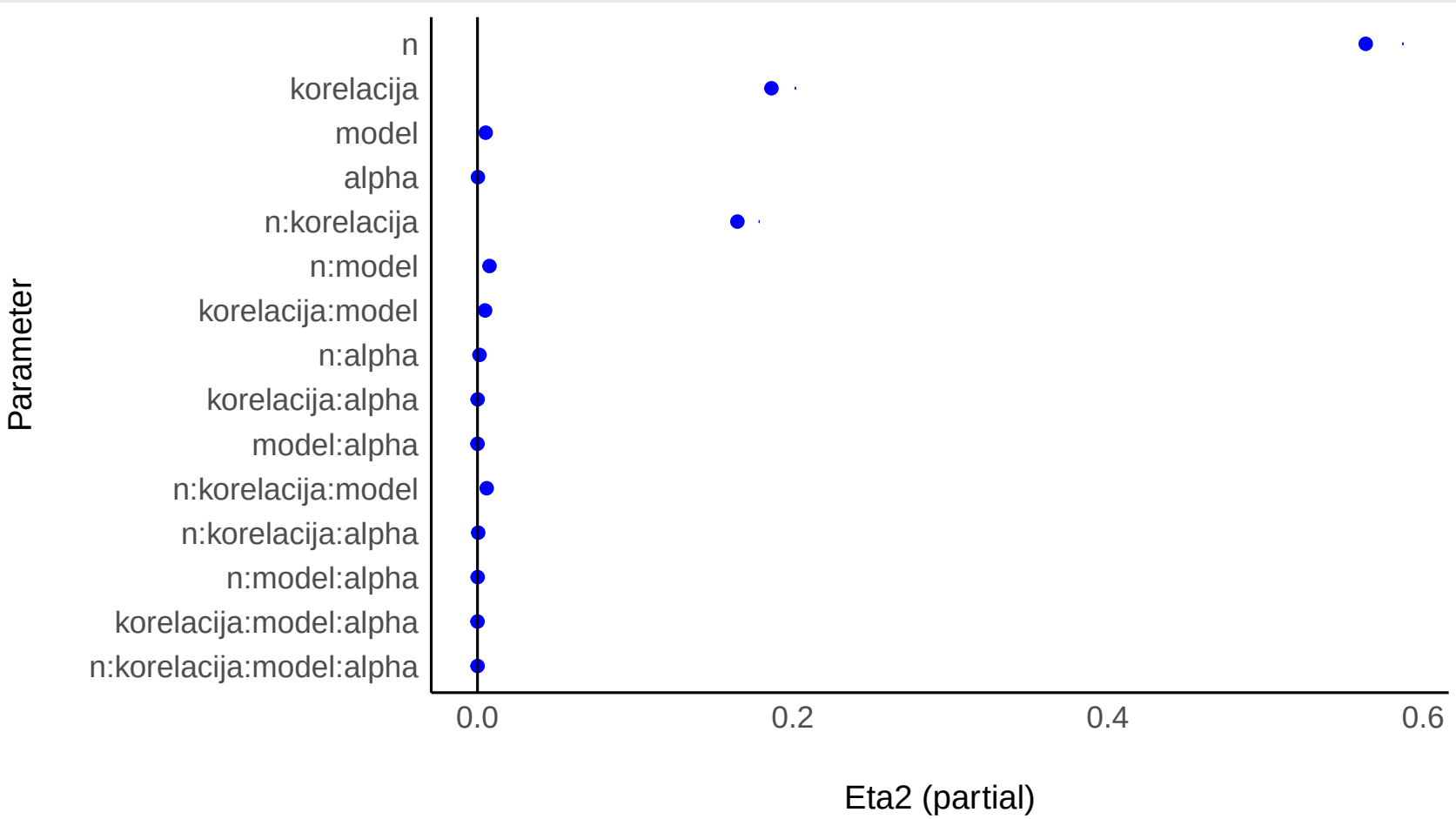
Analiza variance

Pokritost



Analiza variance

Širina IZ



Dopolnitev in izboljšave

- Preverjanje vpliva transformacije (link funkcija v `glm`): Za neko funkcijo g (npr. `g=log()`) je razlika med

$$g(E(Y_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

$$E(g(Y_i)) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

- Skaliranje variance pojasnjevalnih spremenljivk
- Različni načini ocenjevanja IZ

Zaključek

Hvala za pozornost!