Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2021-01-23

Kazalo vsebine

Uvod

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. Ti dejavniki so velikost vzorca, moč korelacije med pojasnjevalnimi spremenljivkami, asimetrija porazdelitve ostankov, asimetrija porazdelitve pojasnjevalnih spremenljivk ter število vključenih spremenljivk v modelu. Poleg tega bomo preverjali, če probleme kršenja predpostavk lahko (vsaj delno) odpravimo z uporabo posplošenih linearnih modelov. Ob koncu naloge bomo preverili še, kako dobro se *link* funkcija v posplošenih modelih obnese v primeru logaritemske transoformacije pojasnjevalne spremenljivke.

Teoretični del

Naloga je osredotočena na linearno regresijo, ki spada pod posplošene linearne modele. V tem poglavju so najprej bolj splošno predstavljeni posplošeni linearni modeli, nato pa podrobneje model linearne regresije in metode, s pomočjo katerih lahko ocenjujemo regresijske koeficiente.

Posplošeni linearni modeli

Posplošeni linearni mešani model izrazimo kot

$$Y = X\beta + Z\alpha + \epsilon,$$

kjer je Y opazovani slučajni vektor, X matrika znanih vrednosti pojasnjevalnih spremenljivk, β neznan vektor regresijskih koeficientov (fiksni učinki), Z znana matrika, α vektor naključnih učinkov in ϵ vektor napak. α in ϵ sta neopazovana. Predpostavimo, da sta nekorelirana.

V matrični obliki model izgleda takole:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & 0 & \dots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} Z_{1,1} & Z_{1,2} & \dots & Z_{1,q} \\ Z_{2,1} & Z_{2,2} & \dots & Z_{2,q} \\ \vdots & \vdots & & \vdots \\ Z_{n,1} & Z_{n,2} & \dots & Z_{n,q} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_q \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Linearni mešani modeli se delijo na Gaussove ali normalne in ne-Gaussove. Pomembna predpostavka pri normalnih linearnih mešanih modelih je normalna porazdeljenost vektorja slučajnih učinkov $\alpha \sim N(0, \sigma^2 I_q)$ in vektorja slučajnih odstopanj $\epsilon \sim N(0, \tau^2 I_n)$, ki nista nujno enakih razsežnosti. Druga pomembna predpostavka je neodvisnost slučajnih vektorjev α in ϵ. Prednost uporabe nenormalnih linearnih mešanih modelov pred normalnimi je v tem, da so bolj fleksibilni za modeliranje (Maver, 2018, str. 6).

Linearna regresija

Linearna regresija je statistični model, ki ga v najbolj enostavni obliki lahko zapišemo kot:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

kjer so ϵ_i med seboj neodvisne slučajne spremenljivke, x_i pa dane vrednosti. Velja $\epsilon_i \sim N(0, \sigma^2)$ za vsak i in tako $Y_i \sim N(\beta_0 + \beta_1 x_i,\ \sigma^2)$. Model lahko razširimo na več linearnih parametrov:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

kjer so ϵ_i neodvisne enako porazdeljene slučajne spremenljivke, za $1 \leq i \leq n$.

Lahko ga zapišemo tudi v matrični obliki:

$$Y = X\beta + \epsilon.$$

Med temeljne predpostavke regresijskega modela spada predpostavka, da med neodvisnimi spremenljivkami ni popolne kolinearnosti ali multikolinearnosti. Najbolj tipičen vzrok za kršenje te predpostavke je, da smo v model kot neodvisni vključili dve spremenljivki, med katerima obstaja močna linearna povezanost. Do multikolinearnosti pride tudi, če v model vključimo več spremenljivk kot je velikost vzorca. Na multikolinearnost posumimo, če se v modelu determinacijski koeficient izkaže za statistično značilnega, od regresijskih koeficientov pa nobeden.

Opazovanja so med seboj neodvisna. V primeru kršenja te predpostavke je smiselno uporabiti posplošene linearne modele, običajno longitudinalni (vzdolžni) model. Vse predpostavke linearnega regresijskega modela so navedene v naslednjem razdelku.

Metoda najmanjših kvadratov (MNK)

Pri 16 letih jo je odkril nemški matematik Carl F. Gauss. Zaradi svojih lastnosti je najbolj razširjena metoda ocenjevanja regresijskih koeficientov (Pfajfar, 2018, str.53).

Pri MNK na primeru osnovnega regresijskega modela velikosti $p = 1$ iščemo β_0 in $\beta 1$ tako, da bo vsota kvadratov ostankov najmanjša možna. Pri danih (x_i, y_i) torej iščemo

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Za razumevanje oznak v predpostavkah metode ločimo dva modela, in sicer linearni vzorčni regresijski model $y_i = b_1 + b_2 x_i + e_i$ in linearni populacijski regresijski model $y = \beta_1 + \beta_2 x_i + u_i$. Pfajfar (2018) navaja naslednje predpostavke metode najmanjših kvadratov:

- 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 $i \neq j$
- nekoreliranost med pojasnjevalnimi spremenljivkami in slučajno spremenljivko u: $Cov(x_2, u) = Cov(x_3, u) = ... = Cov(x_k, u) = 0$
- število opazovanj mora presegati število ocenjenih parametrov oz. pojasnjevalnih spremenljivk: $n > k$
- $Var(X)$ je končno pozitivno število
- pravilno specificiran regresijski model: vključene vse relevantne pojasnjevalne spremenljivke in izbrana ustrezna funkcijska oblika modela
- odsotnost popolne multikolinearnosti: $\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0$
- slučajna spremenljivka u je normalno porazdeljena: $u_i \sim N(0, \sigma_u^2)$. Posledično je pogojna porazdelitev odvisne spremenljivke y tudi normalna in sicer $N(\beta_1 x_{1i} + ... + \beta_k x_{ki}, \sigma_u^2)$

Metoda iterativnega uteženega povprečja najmanjših kvadratov (IWLS)

Naj bo Y vektor meritev in X matrika znanih konstant. Naj bo $E(Y) = X\beta$, kjer β kot do sedaj predstavlja vektor neznanih regresijskih koeficientov. Cenilko za β se po uteženi metodi najmanjših kvadratov dobi z minimizacijo izraza

$$(Y - X\beta)' W (Y - X\beta), \tag{1}$$

kjer je W znana simetrična matrika uteži.

Brez škode za splošnost naj bo rang matrike X poln in naj velja rang$X = p$. Potem je za vsako nesingularno (simetrično) matriko W minimum izraza (1) enak

$$\hat{\beta}_W = (X'WX)^{-1} X'WY. \tag{2}$$

Cenilko za β po običajni metodi najmanjših kvadratov (ang. ordinary least squares, OLS) se dobi kot poseben primer, za $W = I$:

$$\hat{\beta}_I = (X'X)^{-1}X'Y. \tag{3}$$

Izkaže se, da je v smislu čim manjše variance optimalna izbira za matriko W matrika $W = V^{-1}$, kjer je $V = Var(Y)$. Tako dobljena cenilka za parameter β je najboljša, saj je z njeno uporabo dosežena najmanjša možna variabilnost med vsemi drugimi alternativami. V tem primeru se dobljeni cenilki za β reče najboljša linearna nepristranska cenilka ali $BLUE$ (ang. best linear unbiased estimator):

$$\hat{\beta}_{BLUE} = (X'V^{-1}X)^{-1}X'V^{-1}Y. \tag{4}$$

V enačbi za β_{BLUE} nastopa tudi V, ki pa tipično ni znana. Zaradi poenostavitve je v nadaljevanju prikazan postopek izračuna cenilke $BLUE$ zgolj na uravnoteženem primeru. Naj bo $Y_{ij}, j = 1, ..., \tilde{m}$, vektor meritev na $i-$tem posamezniku, kjer je $\tilde{m}$ fiksno število. V uravnoteženem primeru so na vseh posameznikih meritve pridobljene ob določenih časovnih trenutkih $t_1, ..., t_{\tilde{m}}$. Za $i-$tega posameznika se lahko vektor meritev zapiše kot $Y_i = (Y_{ij})_{\leq j \leq \tilde{m}}, i = 1, ..., n$. Naj bodo $Y_1, ..., Y_n$ med seboj neodvisni in naj za njih velja $E(Y_i) = X_i\beta$ in $Var(Y_i) = V_0$. Tu je X_i matrika znanih konstant in $V_0 = (v_{qr})_{1\leq q,r\leq \tilde{m}}$ neznana variančno kovariančna matrika. Iz tega sledi, da je $V = diag(V_0, ..., V_0)$. Ker je število meritev $\tilde{m}$ na vsakem posamezniku fiksno, je mogoče poiskati dosledno cenilko za V. Če bi bil parameter β znan, bi bila dosledna cenilka za V kar

$$\hat{V} = diag(\hat{V}_0, ..., \hat{V}_0),$$

kjer je

$$\hat{V}_0 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta)(Y_i - X_i\beta)'. \tag{5}$$

Če bi bila V znana, bi lahko za izračun najboljše linearne nepristranske cenilke za β uporabili (4), če pa bi poznali β, bi z (5) dobili dosledno cenilko za V.

Metodi, kjer ni treba poznati ne β, ne V, pa se reče iterativno uteženo povprečje najmanjših kvadratov (ang. iterative weighted least squares, IWLS). Postopek omenjene metode je sledeč:

- Najprej se izračuna cenilka za β po običajni metodi najmanjših kvadratov s pomočjo (3).
- Nato se izračuna $\hat{V}$ po (5), kjer je β zamenjan z $\hat{\beta}_I$ izračunanim en korak prej.
- V zadnjem koraku pa se na desni strani (4) matriko V zamenja z njeno cenilko $\hat{V}$, izračunano na prejšnjem koraku.

Na tak način se dobi cenilka za β po prvi iteraciji, nato pa se postopek ponavlja. Pod predpostavko normalnosti se izkaže, da če IWLS konvergira, bo cenilka v limiti enaka cenilki, dobljeni po metodi največjega verjetja (celotno podpoglavje je povzeto po Maver, 2018, strani 19-21).

Generiranje podatkov

Parametri

Fiksni parametri pri generiranju podatkov so sledeči:

- formula za generiranje podatkov:

$$y_i = 1 + x_{1i} + x_{2i} + 0x_{3i} + \epsilon_i.$$

Pri generiranju podatkov bomo spreminjali sledeče:

- velikost vzorca $n \in \{10, 50, 100, 500, 1000\}$;
- korelacija med pojasnjevalnimi spremenljivkami ($cor \in \{0, 0.3, 0.6, 0.9\}$);

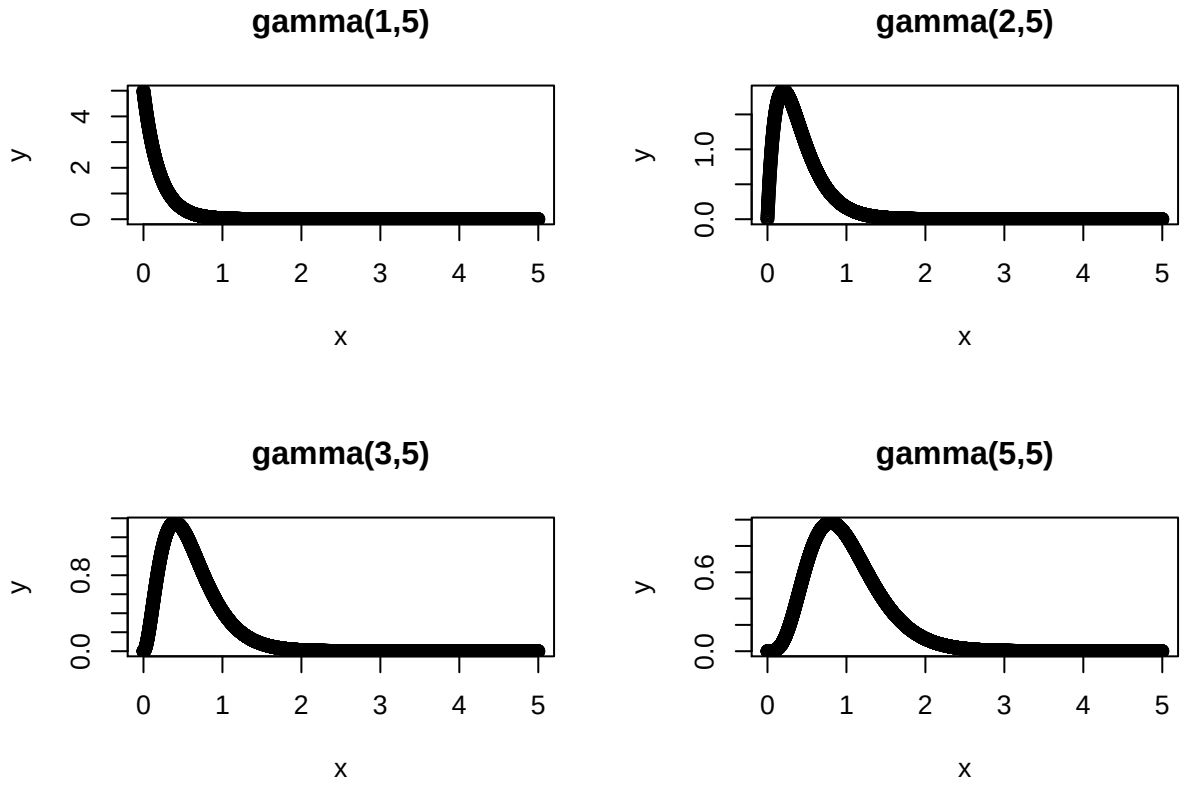
- porazdelitev pojasnjevalnih spremenljivk: $X_j \sim Gamma(\delta, 5)$, $j = 1, 2, 3$, $\delta = 2, 5$;
- porazdelitev napak ($Gamma(\alpha, 5)$), kjer bomo parameter α spreminjali tako, da dobimo različno močno asimetrične porazdelitve ($\alpha \in \{1, 3, 5\}$;
- v modelu ne upoštevamo vseh neodvisnih spremenljivk (spreminjamo število spremenljivk, ki jih upoštevamo): enkrat vključimo vse spremenljivke, enkrat izločimo X_3 (ki nima vpliva na odzivno spremenljivko), enkrat pa izločimo X_2.

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

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

Pri pregledu literature sva ugotovili, da za generiranje odvisnih gama spremenljivk lahko uporabimo kar funkcijo *rmvgamma()* iz paketa `lcmix` in si s tem olajšamo delo pri generiranju podatkov. Funkcija sprejme naslednje parametre: n (število vektorjev, ki jih želimo generirati), *corr* (korelacijska matrika) ter parametra `shape` in `rate`, ki sta privzeto nastavljena na 1. Kot rezultat dobimo matriko z n vrsticami in $ncol(corr)$ stolpci, v kateri so vrednosti koreliranih gamma spremenljivk (mvgama, 2020). Koda, na podlagi katere funkcija generira vrednosti, je predstavljena v viru lcmix, 2021.

Prikaz porazdelitev, na podlagi katerih so generirane pojasnjevalne spremenljivke in ostanki, je predstavljen na naslednji sliki.



Slika 1: Različne porazdelitve gama

V poglavju `Transformacija odzivne spremenljivke` obravnavamo poseben primer, kjer za generiranje

podatkov uporabimo sledečo formulo:

$$y_i = exp(1 + x_{1i} + x_{2i} + 0x_{3i} + \epsilon_i)$$

Enako kot prej imamo 5 različnih velikosti vzorcev, 4 različne korelacije med pojasnjevalnimi spremenljivkami, 3 porazdelitve ostankov, 2 porazdelitvi pojasnjevalnih spremenljivk in 2 metodi, ne obravnavamo pa modelov z manjkajočimi spremenljivkami. Model *lm* oblikujemo po sledeči formuli:

$$log(Y) \sim \beta X + \epsilon,$$

v modelu *glm* pa uporabimo kar

$$Y \sim \beta X + \epsilon$$

in za upoštevanje logaritemske transformacije uporabimo `family=gamma(link="log")` (podrobneje razloženo v naslednjem razdelku). Zanima nas predvsem razlika med tema dvema pristopoma.

Funkciji lm() in glm()

Funkcija *lm()* se uporablja za ocenjevanje linearnih modelov. Avtomatično uporablja osnovno metodo najmanjših kvadratov, lahko pa nastavimo tudi na metodo uteženih najmanjših kvadratov (lm iz RDocumentation, 2020). V tej seminarski nalogi uporabljamo samo osnovno metodo najmanjših kvadratov.

V linearnem modelu $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i$ velja $E(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$. Slednjo enačbo lahko z uporabo primerno definirane funkcije g posplošimo do

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

Tu z indeksom i označujemo i-tega posameznika. Prejšnja enačba je poseben primer, kjer je g identiteta. Z uporabo funkcije *glm()* in znotraj primerno definirane funkcije g lahko generiramo več posplošenih linearnih modelov. Prednost te funkcije je tudi v tem, da lahko poleg normalne porazdelitve nastavimo še katero drugo porazdelitev ostankov. To naredimo tako, da npr. v primeru gama porazdelitve ostankov znotraj funkcije *glm()* definiramo *family=Gamma(link="identity")*. Tu *identity* pomeni, da za funkcijo g vzamemo kar identiteto, možne izbire pa so še *log*, *inverse*, *logit* in druge (možna je tudi nova definicija). Parameter *family* določa porazdelitev ostankov. Na voljo imamo normalno, binomsko, gama, inverzno normalno, Poissonovo, kvazi (poleg *link* funkcije nastavimo še varianco), kvazi-binomsko in kvazi-Poissonovo. Funkcija *glm* parametre modela ocenjuje po metodi iterativnega uteženega povprečja najmanjših kvadratov (glm iz RDocumentation, 2020).

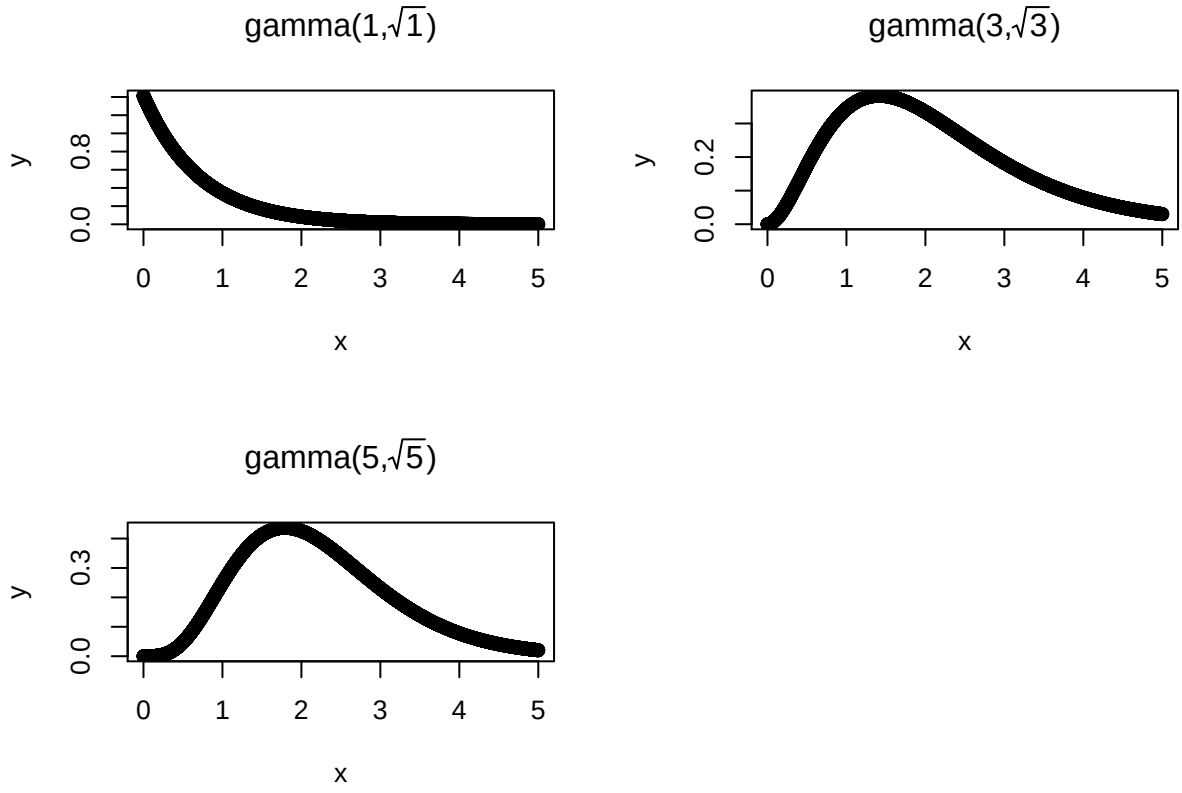
Ocenjevanje intervalov zaupanja

Pri izračunih intervalov zaupanja uporabimo funkciji *confint* in *confint.default*. Prva predpostavlja normalnost, druga pa temelji na asimptotski normalnosti in računa intervale zaupanja na podlagi standardnih napak. Funkcija *confint* je posebej prilagojena funkcijama *lm* in *glm* - če uporabimo model, ocenjen na enega od teh dveh načinov, funkcija avtomatično pokliče *confint.lm* oziroma *confint.glm* (confint iz RDocumentation, 2020). Iz neznanih razlogov se pri posplošenih linearnih modelih občasno pojavljajo težave, zato v primeru "error" uporabimo *confint.default*. Če pri generiranju podatkov ne skaliramo variance, ni težav z uporabo funkcije *confint*, vendar pa je v tem primeru potrebno paziti pri interpretaciji rezultatov. V kodi za simulacije je nastavljena zanka, ki v primeru napake uporabi funkcijo *confint.default* in neuspeh funkcije *confint* zabeleži pod številko 0 v stolpcu "confint_success" (če ni težav, zabeleži vrednost 1).

Skaliranje variance

Ker zaradi spreminjanja parametrov v gama porazdelitvi ne spreminjamo samo asimetričnosti, ampak tudi varianco in ker vemo, da ima velikost variance vpliv na model, smo spremenljivke napak skalirali. Napake smo skalirali tako, da smo vrednosti delili s teoretično standardno napako in tako poskrbeli, da imajo vse porazdelitve napak enako varianco.

Podrobnejši pregled gama porazdelitve z vsemi dokazi je pripravil K. Siegrist (2020). Če je spremenljivka X porazdeljena po $Gamma(\alpha, \beta)$, je njena pričakovana vrednost enaka $\frac{\alpha}{\beta}$, varianca pa $\frac{\alpha}{\beta^2}$. Za to spremenljivko in neko konstanto c velja $cX \sim Gamma(\alpha, \frac{\beta}{c})$. To lastnost uporabimo pri skaliranju variance in sicer tako, da vrednosti delimo s $\sqrt{\frac{\alpha}{\beta^2}} = \frac{\sqrt{\alpha}}{\beta}$. Tako dobimo skalirano spremenljivko $\frac{\beta}{\sqrt{\alpha}}X \sim Gamma(\alpha, \sqrt{\alpha})$.



Slika 2: Skalirane porazdelitve ostankov

V poglavju `Transformacija odzivne spremenljivke` skaliranja variance nismo uporabili. Razlog je v tem, da program občasno vrača opozorila `task failed - "NA/NaN/Inf in 'x'"`. Ker nas v resnici zanima le razlika med pristopoma *lm* in *glm*, nas različna varianca ostankov pri interpretaciji ne zmoti.

Pričakovanja

Pri večji korelaciji med pojasnjevalnimi spremenljivkami pričakujemo širše intervale zaupanja regresijskih koeficientov ne glede na izbiro metode. Zaradi večje širine intervalov zaupanja ne pričakujemo večjih sprememb v pokritosti.

Večje razlike med metodami pričakujemo predvsem pri manjših velikostih vzorcev in večji asimetriji porazdelitve ostankov. Pri dovolj velikih vzorcih pričakujemo podobne rezultate obeh metod, prav tako pa seveda tudi manjšo variabilnost rezultatov.

Pričakujemo, da lahko kršenje predpostavke o normalni porazdeljenosti ostankov rešimo z uporabo posplošenih linearnih modelov z ustrezno definirano porazdelitvijo ostankov oz. odzivne spremenljivke. Pričakujemo, da bolj kot bo porazdelitev ostankov asimetrična (manjša vrednost parametra α), slabši bodo rezultati funkcije *lm()* in posledično večje razlike med rezultati funkcij *lm()* in *glm()*.

V primeru, ko iz modela izločimo spremenljivko X_3 in nimamo velikih korelacij med pojasnjevalnimi spremenljivkami, ne pričakujemo posebnih sprememb v rezultatih, saj spremenljivka nima vpliva na vrednost pojasnjevalne spremenljivke. Pravzaprav lahko model brez te spremenljivke označimo kot "pravi" model. V primeru visoke korelacije med pojasnjevalnimi spremenljivkami v polnem modelu pričakujemo večje napake pri ocenjevanju regresijskih koeficientov (posledično slabšo pokritost ali širše IZ), zato predvidevamo, da se bo model brez vključene spremenljivke X_3 bolje obnesel. V primeru, ko izločimo spremenljivko X_2, pa pričakujemo večje spremembe v rezultatih - širše intervale zaupanja regresijskih koeficientov in slabšo pokritost.

Porazdelitev pojasnjevalnih spremenljivk preverimo za dve porazdelitvi - $Gamma(2, 5)$, ki je precej asimetrična in $Gamma(5, 5)$, ki je zelo podobna normalni porazdelitvi. Zanima nas, če in kako asimetrija pojasnjevalnih spremenljivk vpliva na ocene regresijskih koeficientov. Pričakujemo, da bo v primeru asimetrične porazdelitve prišlo do manjše pokritosti in večje širine intervalov zaupanja.

V primeru transformacij pričakujemo podobne rezultate med metodama *lm* in *glm*, morda boljše pri slednji. Prednost *glm* je v tem, da lahko nastavimo gama porazdelitev ostankov, za *link* funkcijo pa pričakujemo, da se bo obnesla približno tako dobro kot logaritemska transformacija formule v *lm* funkciji.

Predstavitev rezultatov

Izvedli smo 1000 ponovitev simulacij. Na vsakem koraku obravnavamo 4 različne korelacije, 5 različnih velikosti vzorca, 3 različne porazdelitve ostankov, 2 različni porazdelitvi pojasnjevalnih spremenljivk, 2 metodi (*lm* in *glm*) in 3 modele glede na število spremenljivk - iz tega sledi, da dobimo tabelo s 720 000 vrsticami.

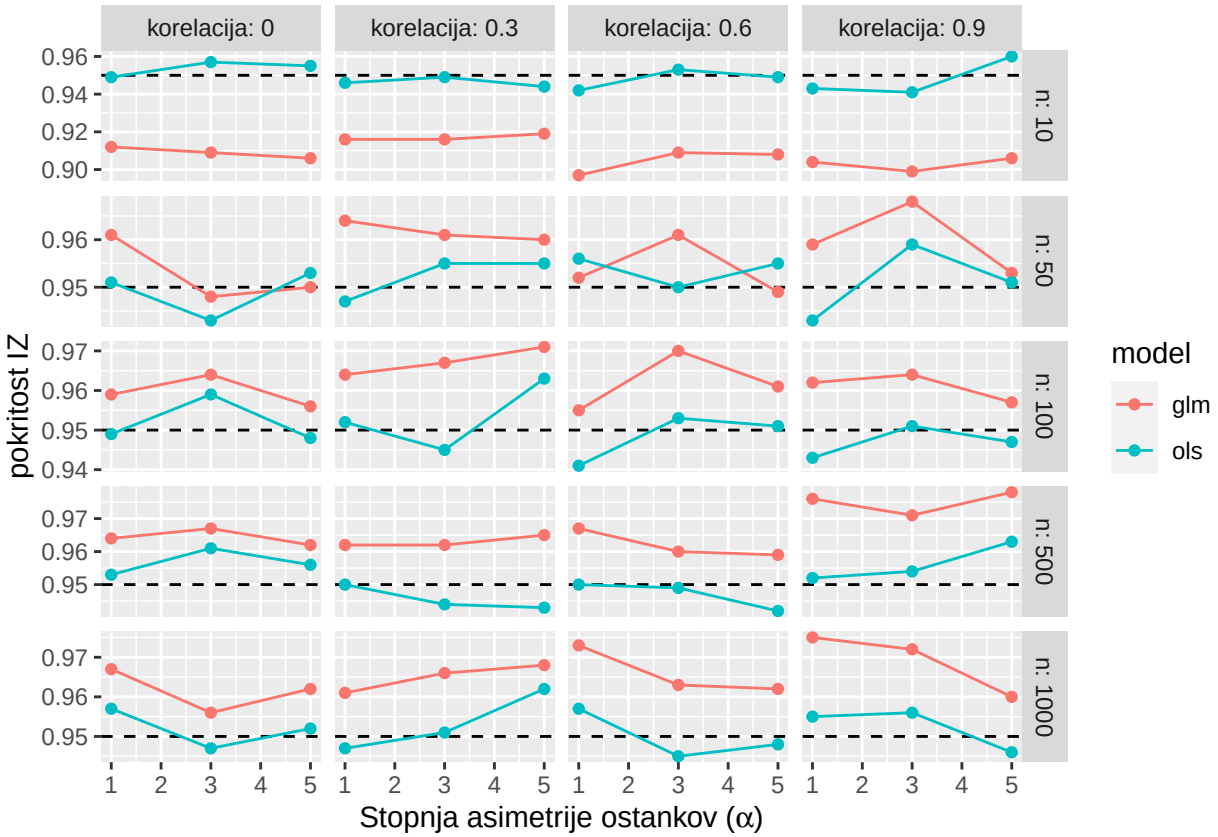
Intervali zaupanja so generirani s pomočjo funkcije *confint()* oziroma v primeru, ko naletimo na `error`, s funkcijo *confint.default()*. Primerov, ko je bilo potrebno uporabiti slednjo funkcijo, je 21, kar znaša 0.003 % vseh ponovitev simulacij.

Rezultati se nanašajo na veliko parametrov in več različnih načinov modeliranja, zato si jih bomo ogledali na različne načine. Ker imata koeficienta X_1 in X_2 enak vpliv na odzivno spremenljivko, X_3 pa nima vpliva, se lahko osredotočimo le na koeficient β_1.

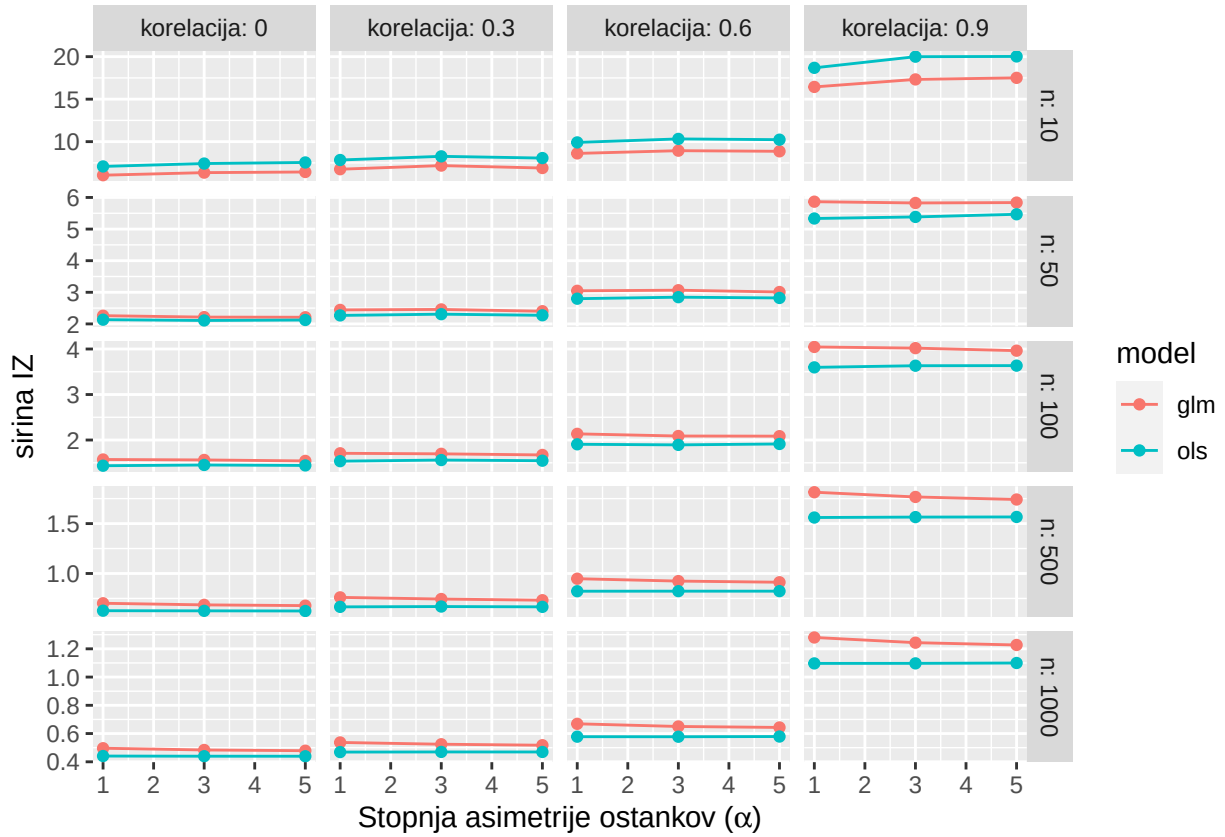
Polni model

Zelo asimetrična porazdelitev pojasnjevalnih spremenljivk

Za začetek si poglejmo rezultate oz. izbrane mere na polnem modelu, kjer so upoštevane vse spremenljivke pri bolj asimetrični porazdelitvi pojasnjevalnih spremenljivk $Gamma(2, 5)$. Naslednje dva grafa tako prikazujeta pokritost in širine intervalov zaupanja za koeficient β_1 za posamezen model (*gls* ali *ols*) glede na velikost vzorca in korelacijo. Črna črtkana črta na grafu pokritja označuje željeno pokritje (0.95).



Slika 3: Pokritost intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 4: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

Iz slik 3 in 4 lahko vidimo, da imata na ocene regresijskih koeficientov precej očiten vpliv pri obeh metodah (*glm* in *ols*) korelacija in velikost vzorca. Z večanjem korelacije se sama pokritost bistveno ne spremeni, se pa močno povečajo širine intervalov zaupanja. Z večanjem velikosti vzorca pridobimo večjo pokritost pri *glm* metodi in ožje intervale zaupanja pri obeh metodah. Pri različnih stopnjah asimetrije ostankov ni opaziti bistvenega vpliva na rezultate. Razberemo lahko, da ima model *glm* pri dovolj velikih vzorcih nekoliko boljšo pokritost. Velja pa tudi, da pri *glm* modelu dobimo nekoliko širše intervale zaupanja, od koder najverjetneje izhaja tudi boljša pokritost. Intervali zaupanja pri metodi *ols* so ožji, posledično pa interval zaupanja za koeficient večkrat ne vsebuje prave vrednosti koeficienta. Razlike v pokritosti sicer niso velike, vseeno pa so prisotne. Zanimivo je dejstvo, da ima samo pri majhnem vzorcu ($n = 10$) metoda *ols* širši interval zaupanja in boljšo pokritost kot *glm*.

Za manjše število izbranih parametrov n in korelacije so rezultati prikazani tudi v spodnjih dveh tabelah, kjer lažje razberemo razlike. Prikazane razlike so v obeh primerih razlike *glm* glede na *ols* metodo, v tabeli s širino intervalov zaupanja so prikazane še razlike v deležu glede na širino intervala zaupanja pri metodi *ols*.

Tabela 1: Pokritost intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	ols	razlika
10	0.3	0.92	0.95	-0.03
10	0.9	0.90	0.95	-0.05
100	0.3	0.97	0.95	0.02
100	0.9	0.96	0.95	0.01

n	korelacija	glm	ols	razlika
1000	0.3	0.96	0.95	0.01
1000	0.9	0.97	0.95	0.02

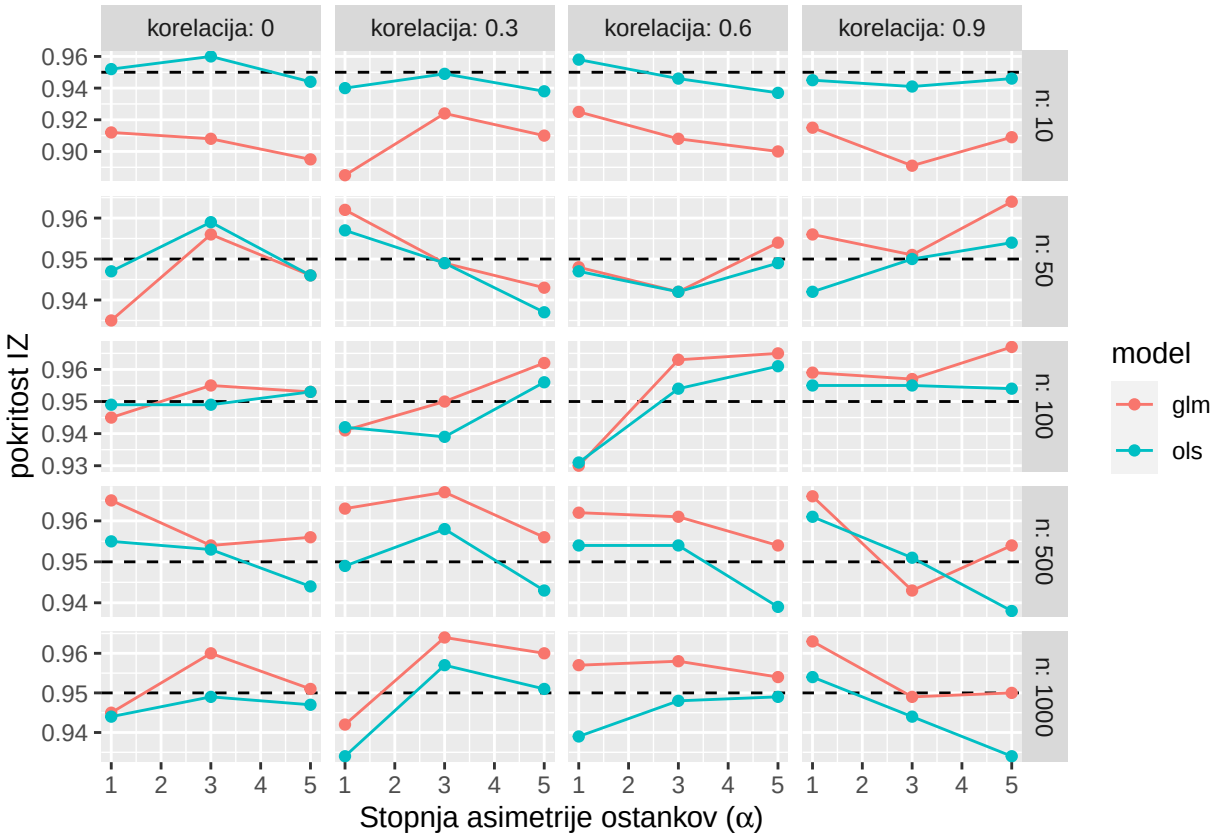
Tabela 2: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	6.94	8.06	-1.12	-0.14
10	0.9	17.09	19.57	-2.48	-0.13
100	0.3	1.69	1.55	0.14	0.09
100	0.9	4.01	3.62	0.39	0.11
1000	0.3	0.53	0.47	0.06	0.13
1000	0.9	1.25	1.10	0.15	0.14

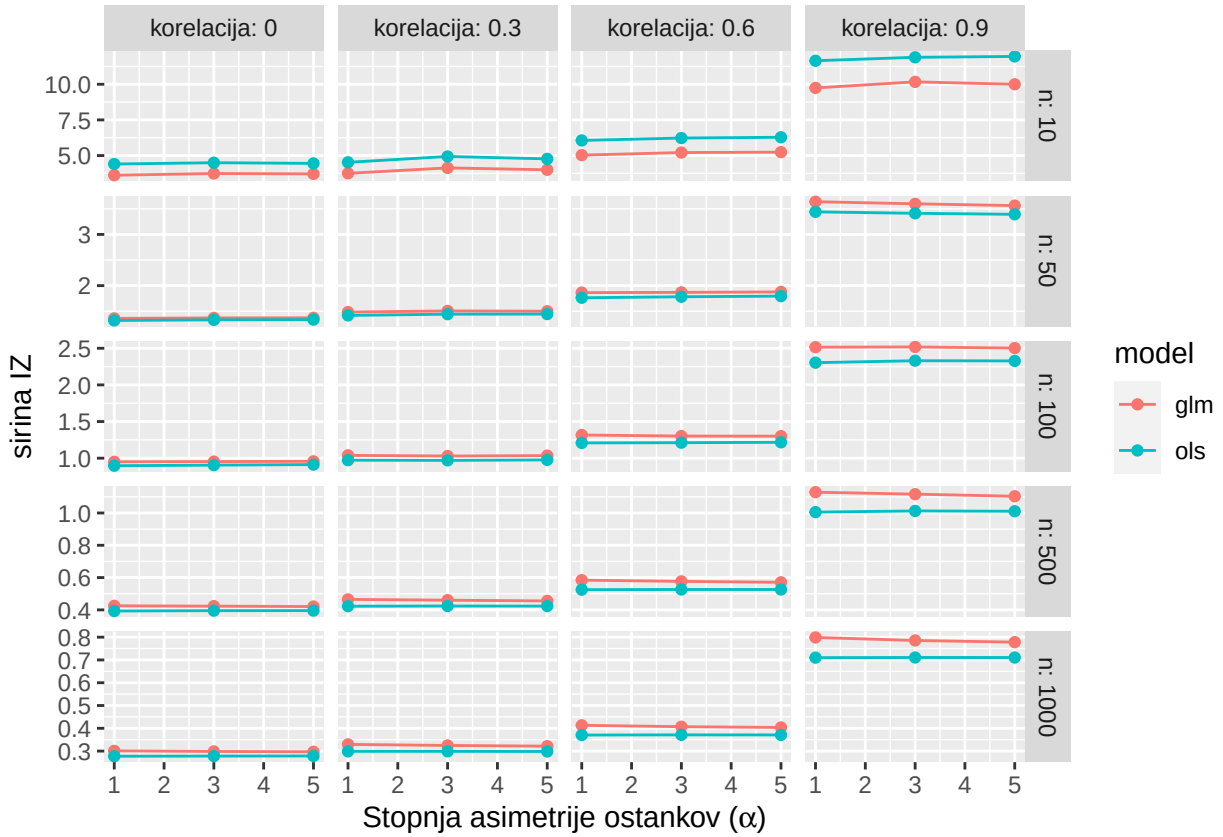
Velja torej, da ima v povprečju *glm* metoda od 1 do 2 ostotni točki boljšo pokritost (z izjemo najmanjšega opazovanega vzorca), vendar pa ima tudi do 14% širše intervale zaupanja. Ta razlika med intervali zaupanja se najbolj poveča z večanjem velikosti vzorca. Visoka korelacija v kombinaciji z majhnim vzorcem je najbolj problematična, saj imamo tam tako manjšo pokritost kot tudi precej široke intervale zaupanja.

Manj asimetrična porazdelitev pojasnjevalnih spremenljivk

Enako kot smo si pogledali za $Gamma(2, 5)$ porazdelitev pojasnjevalnih spremenljivk, poglejmo še za nekoliko bolj simetrično porazdelitev $Gamma(5, 5)$.



Slika 5: Pokritost intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(5, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 6: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

Ponovno hitro opazimo vpliv velikosti vzorca in korelacije ter dejstvo, da sama asimetrija ostankov nima bistvenega vpliva na ocene regresijskih koeficientov. Razlike med metodama so tokrat nekoliko manj očitne, vseeno pa opazimo, da je pri *glm* metodi pokritost pri večjih vzorcih boljša od pokritosti pri *ols* metodi. Kar se tiče širine IZ, so tokrat razlike res majhne z izjemo majhnega vzorca ($n = 10$) in visoke korelacije (0.9). Ponovno ima torej metoda *glm* pri večjih vzorcih boljšo pokritost (a širše intervale zaupanja) kot druga metoda. Za razlike v širinah intervalov zaupanja si moramo pogledati številke v tabelah, kjer bomo lažje razbrali ali so le-te manjše ali večje kot prej. Predvsem lahko opazimo, da če primerjamo graf širine intervalov zaupanja s prejšnjih grafom, kjer je upoštevana asimetrična porazdelitev, so intervali v splošnem nekoliko ožji.

Tabela 3: Pokritost intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	ols	razlika
10	0.3	0.91	0.94	-0.03
10	0.9	0.90	0.94	-0.04
100	0.3	0.95	0.95	0.00
100	0.9	0.96	0.95	0.01
1000	0.3	0.96	0.95	0.01
1000	0.9	0.95	0.94	0.01

Tabela 4: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in $Gamma(5, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	3.96	4.74	-0.78	-0.16
10	0.9	9.97	11.83	-1.86	-0.16
100	0.3	1.03	0.97	0.06	0.06
100	0.9	2.51	2.32	0.19	0.08
1000	0.3	0.33	0.30	0.03	0.10
1000	0.9	0.79	0.71	0.08	0.11

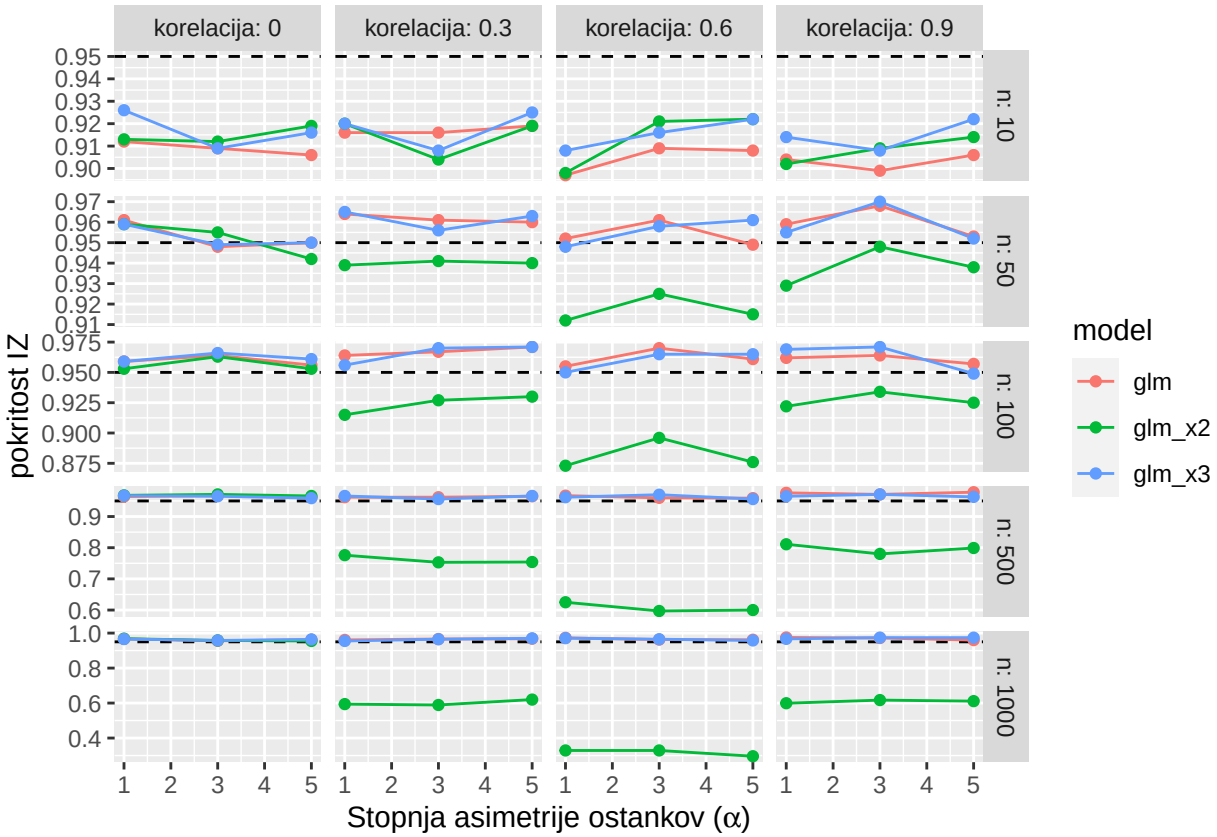
Številke v tabelah potrjujejo naša opažanja. Med metodama `glm` in `ols` je pri pokritosti manjša razlika in sicer je pri dovolj velikih vzorcih tokrat pokritost pri metodi `glm` višja za do 1 o.t. Prav tako je manjša razlika v širini intervalov zaupanja. Ponovno pa velja, da se (procentualna) razlika v širini z večanjem velikosti vzorca povečuje v prid metode *glm*.

Vpliv odstranjevanja spremenljivk

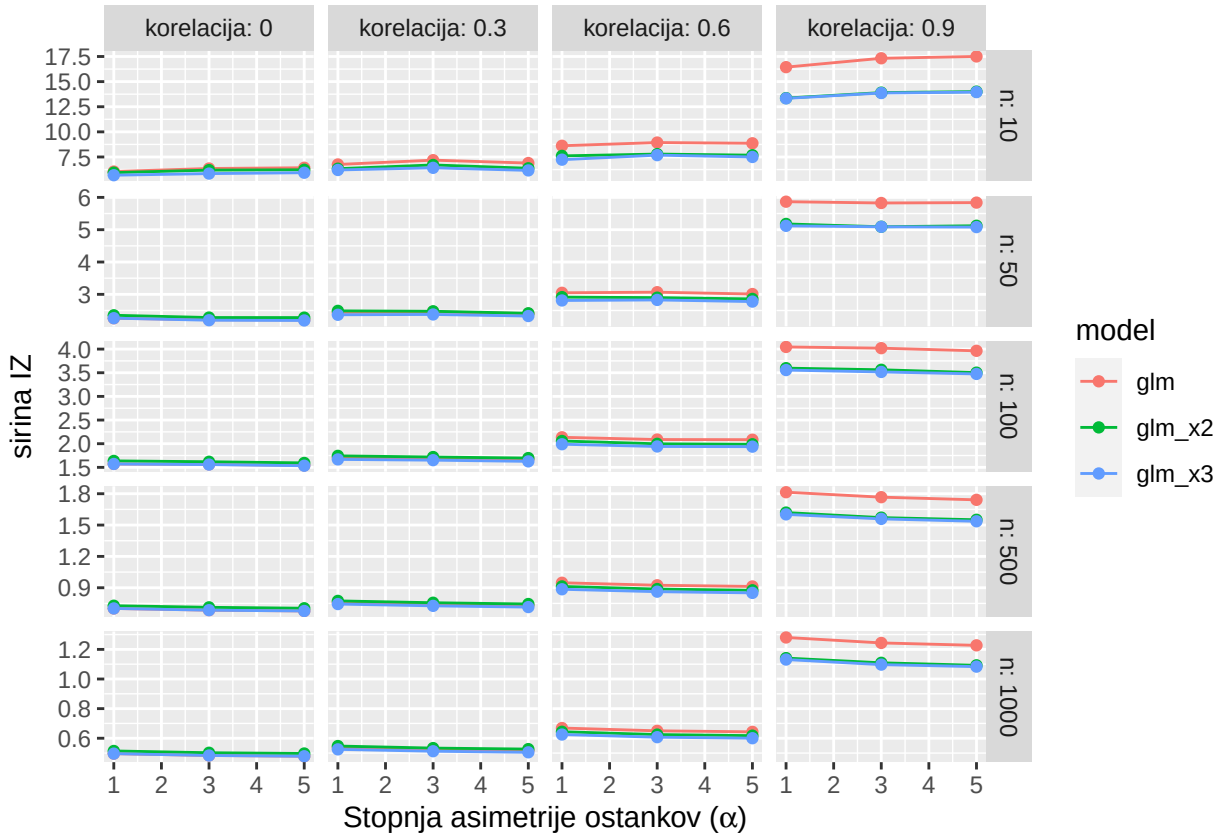
Poglejmo si še vpliv izločanja spremenljivk. Zanima nas, kakšen je ta vpliv, ali se razlikuje med metodama in ali se razlikuje med različno asimetričnima porazdelitvama pojasnjevalnih spremenljivk.

Zelo asimetrična porazdelitev pojasnjevalnih spremenljivk

Za začetek si poglejmo, kako se razlikujejo rezultati pri posamezni metodi, potem pa bomo primerjali še posamezne modele z odstranjenimi spremenljivkami pri obeh metodah hkrati. Zaradi podobnosti med rezultati obeh metod so rezultati modelov (glede na število spremenljivk) predstavljeni le za *glm* metodo.



Slika 7: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 8: Širina intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

Pri majhnem vzorcu ni velikih razlik pri pokritosti glede na model, je pa predvsem ob prisotnosti višje korelacije nekaj razlik v širini intervalov zapanja. Odstranitev spremenljivke X_3 nima večjega vpliva na pokritost, dobimo pa nekoliko ožje intervale zaupanja. Večji vpliv ima odstranitev spremenljivke X_2, ki ima dejanski vpliv na odvisno spremenljivko. V splošnem so intervali zaupanja v primeru izključene spremenljivke pri višji korelaciji ožji kot v polnem modelu, vendar pa ne smemo pozabiti, da je v primeru izključitve X_2 zato slabša pokritost. Kar je zanimivo, je, da se pokritost v primeru modela `glm_x2` močno poslabša, ko povečamo vzorec (vendar le v primeru neničelne korelacije).

Najboljši model dobimo, ko izključimo X_3, ki nima vpliva, saj imamo podobno pokritost kot pri polnem modelu, a ožji interval zaupanja. V naslednjih tabelah, ki potrjujo naša opažanja, si lahko bolje ogledamo dejanske številke.

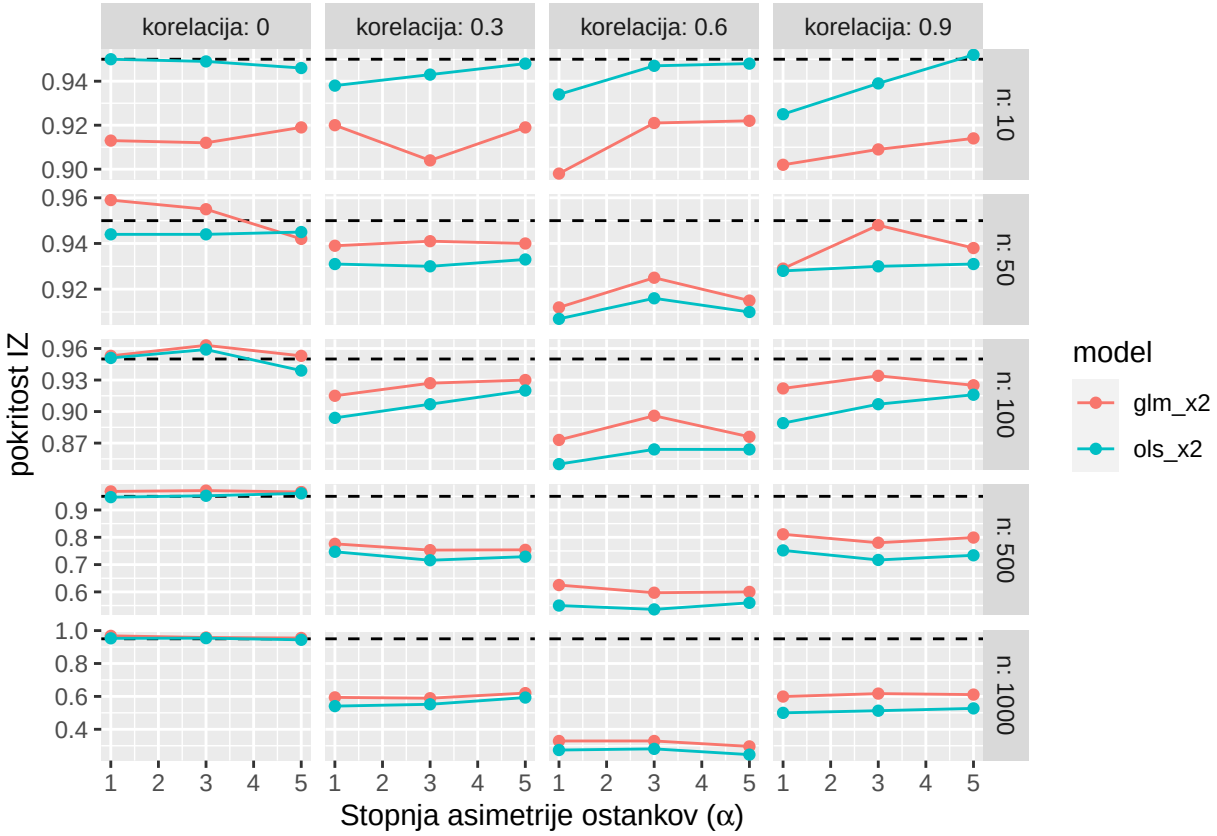
Tabela 5: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	glm_x2	glm_x3
10	0.3	0.92	0.91	0.92
10	0.9	0.90	0.91	0.91
100	0.3	0.97	0.92	0.97
100	0.9	0.96	0.93	0.96
1000	0.3	0.96	0.60	0.96
1000	0.9	0.97	0.61	0.97

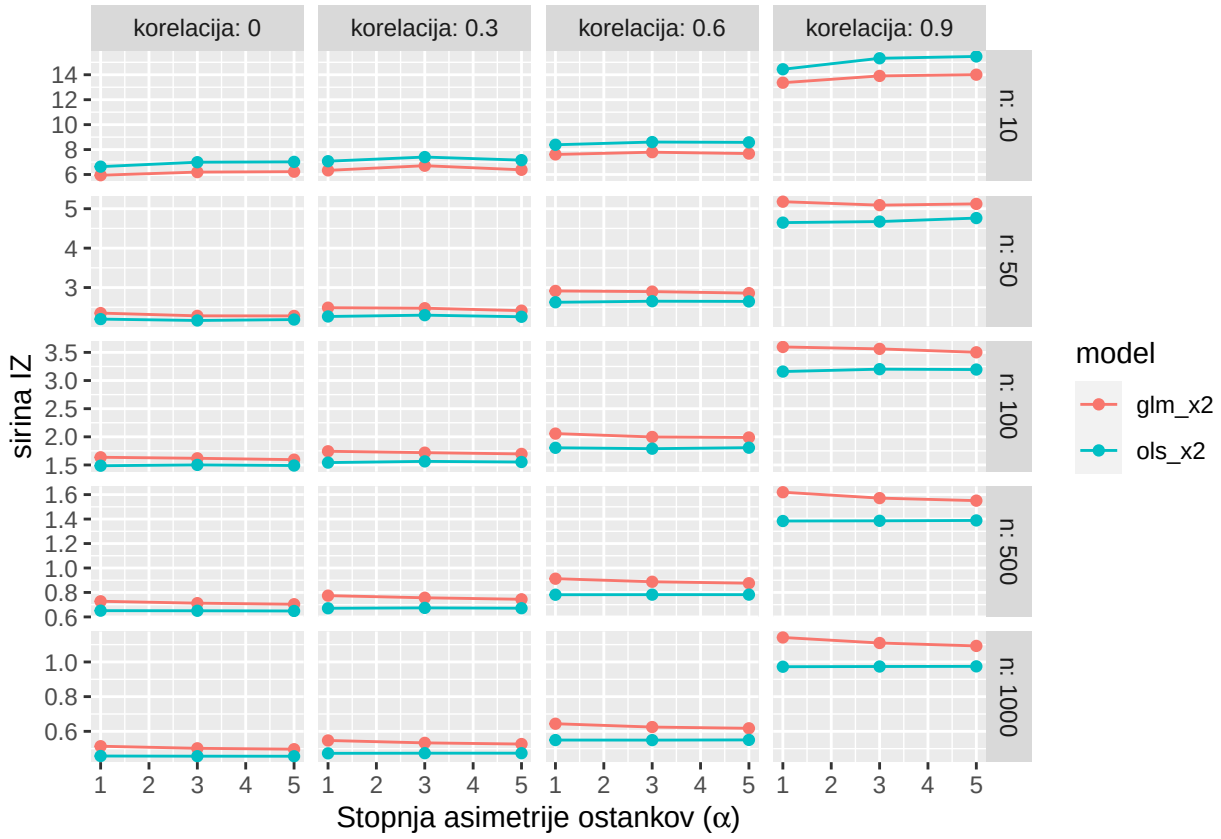
Tabela 6: Širina intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	glm_x2	glm_x3
10	0.3	6.94	6.47	6.26
10	0.9	17.09	13.76	13.72
100	0.3	1.69	1.72	1.65
100	0.9	4.01	3.55	3.52
1000	0.3	0.53	0.54	0.51
1000	0.9	1.25	1.11	1.10

Poglejmo si še razlike med rezultati funkcij *glm* in *lm*, ko v modelu izključimo eno od spremenljivk. Najprej si poglejmo rezultate modela brez X_2.



Slika 9: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 10: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

Podobno kot pri polnih modelih ima metoda *glm* pri dovolj velikih vzorcih boljšo pokritost, a širše intervale zaupanja kot *ols*. Razlika med metodama je bolj očitna pri višji korelaciji. Pri večjih vzorcih glede na polni model nastane večja razlika v pokritosti. Razlika v širini intervala zaupanja je podobna kot v polnem modelu. Sledita še tabeli, v katerih so bolje razvidne številske razlike med metodama.

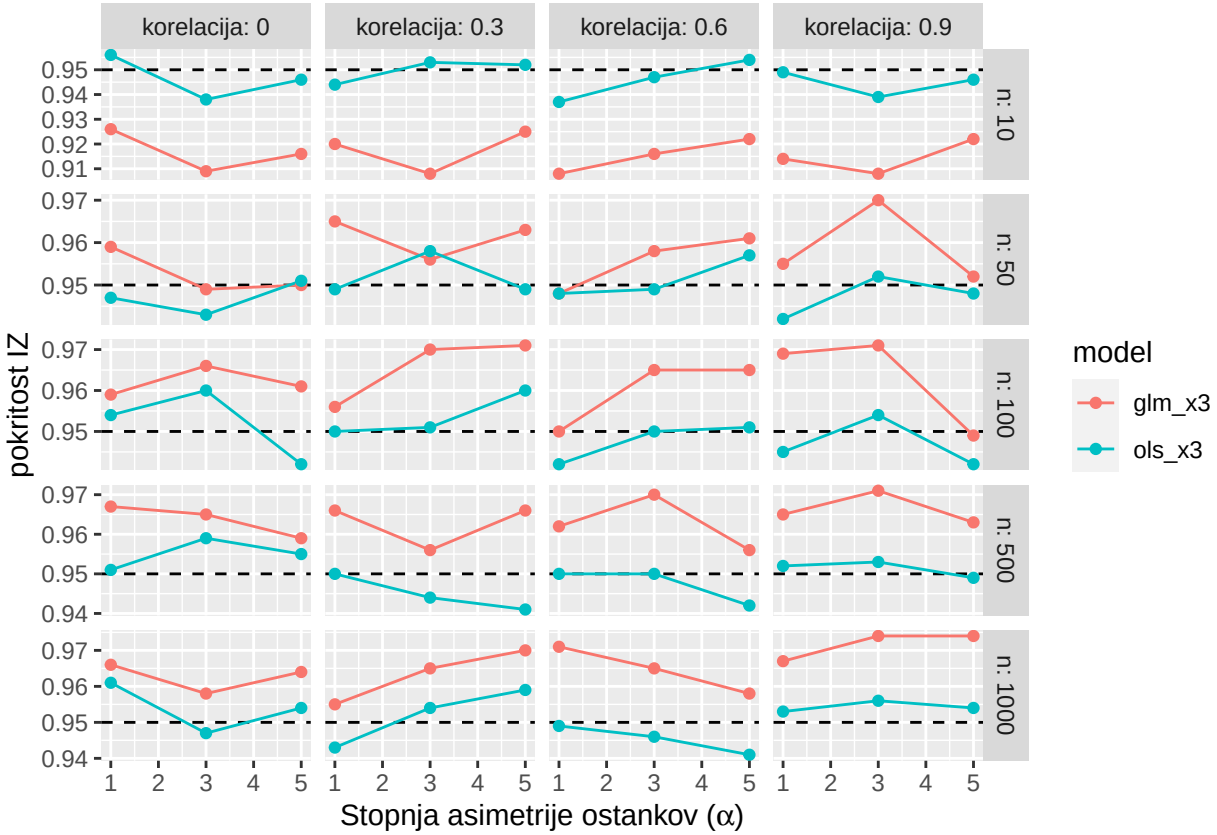
Tabela 7: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika
10	0.3	0.91	0.94	-0.03
10	0.9	0.91	0.94	-0.03
100	0.3	0.92	0.91	0.01
100	0.9	0.93	0.90	0.03
1000	0.3	0.60	0.56	0.04
1000	0.9	0.61	0.51	0.10

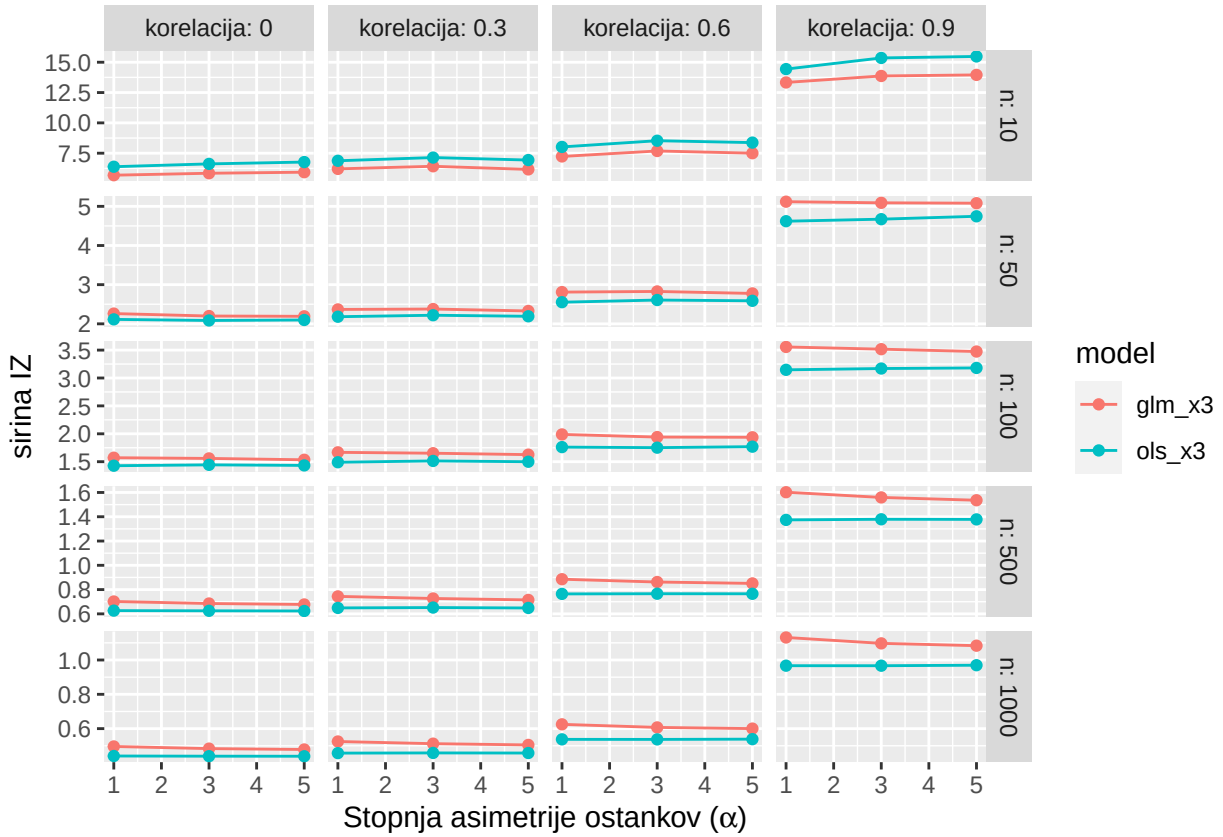
Tabela 8: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika	razlika_pct
10	0.3	6.47	7.21	-0.74	-0.10
10	0.9	13.76	15.08	-1.32	-0.09
100	0.3	1.72	1.55	0.17	0.11
100	0.9	3.55	3.19	0.36	0.11
1000	0.3	0.54	0.47	0.07	0.15
1000	0.9	1.11	0.97	0.14	0.14

Kaj pa, če odstranimo X_3?



Slika 11: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 12: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(2,5)$ porazdelitvi pojasnjevalnih spremenljivk

Rezultati so zelo podobni kot v primeru polnega modela. Model *glm* ima z izjemo majhnega vzorca ($n = 10$) pokritost celo boljšo od željene, intervali zaupanja pa so v primerjavi z *ols* nekoliko širši. V primeru zelo majhnega vzorca je očitno bolje izbrati *ols* model, saj ima kljub nekoliko širšim IZ za 3 o.t. boljšo pokritost.

Tabela 9: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(2,5)$ porazdelitvi pojasnjevalnih spremenljivk

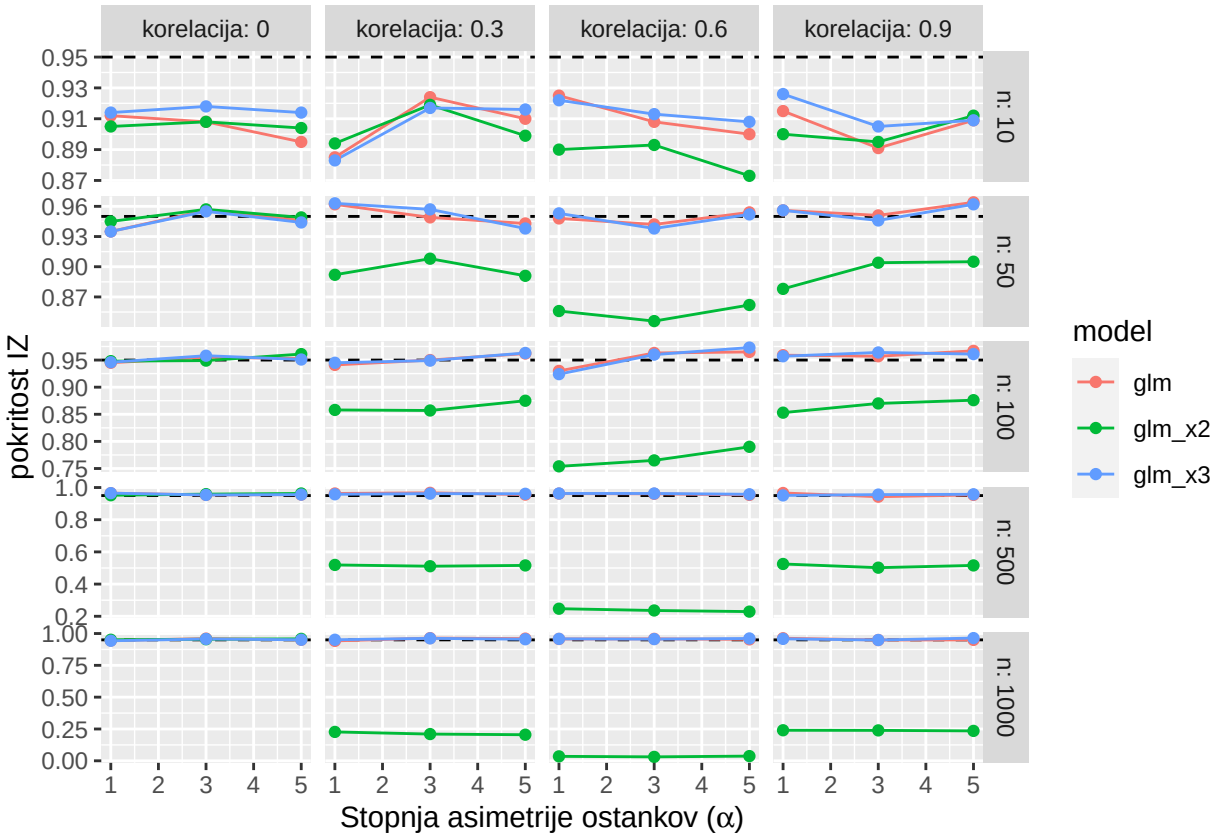
n	korelacija	glm_x3	ols_x3	razlika
10	0.3	0.92	0.95	-0.03
10	0.9	0.91	0.94	-0.03
100	0.3	0.97	0.95	0.02
100	0.9	0.96	0.95	0.01
1000	0.3	0.96	0.95	0.01
1000	0.9	0.97	0.95	0.02

Tabela 10: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(2,5)$ porazdelitvi pojasnjevalnih spremenljivk

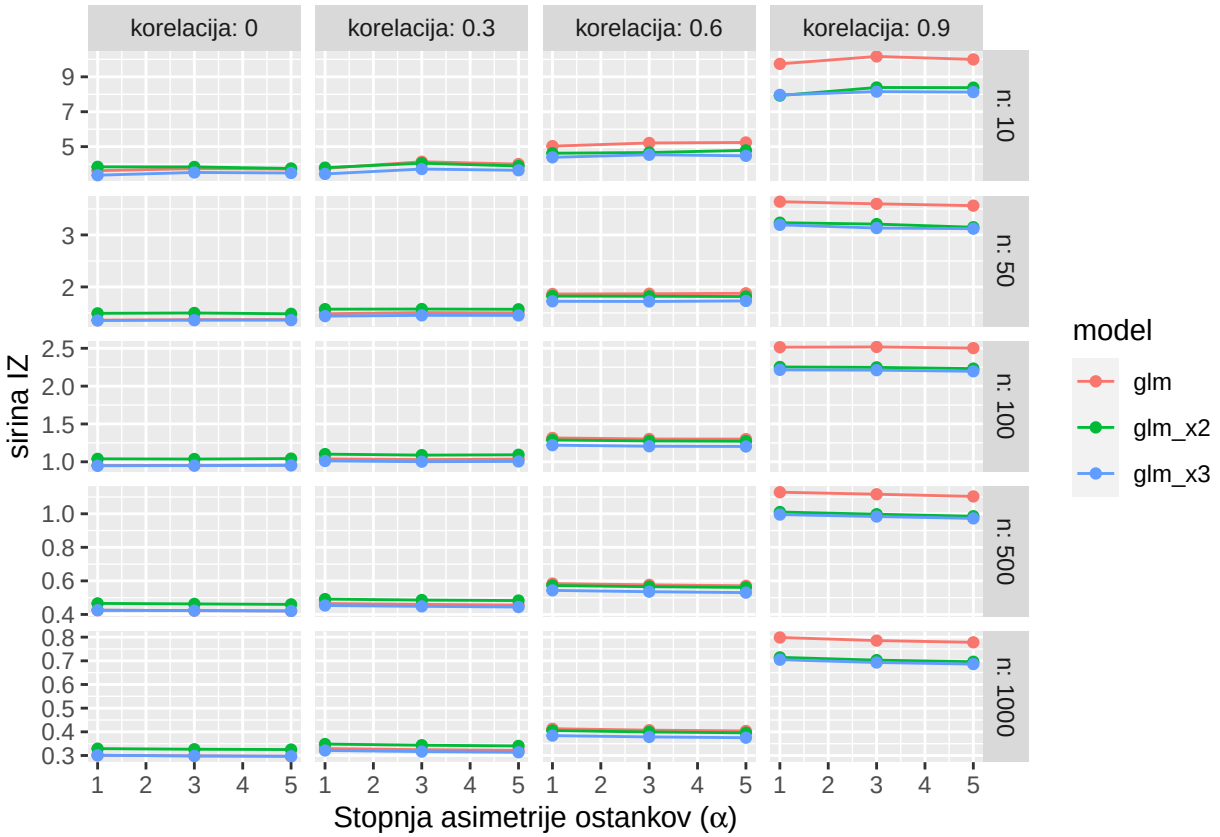
n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
10	0.3	6.26	6.98	-0.72	-0.10
10	0.9	13.72	15.09	-1.37	-0.09

n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
100	0.3	1.65	1.50	0.15	0.10
100	0.9	3.52	3.17	0.35	0.11
1000	0.3	0.51	0.46	0.05	0.11
1000	0.9	1.10	0.97	0.13	0.13

Manj asimetrična porazdelitev pojasnjevalnih spremenljivk



Slika 13: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(5, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 14: Širina intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

Rezultati so na videz podobni tistim z bolj asimetrično porazdelitvijo pojasnjevalnih spremenljivk. Vplivi ostalih parametrov, ki jih spreminjamo, ostajajo enaki, smo pa z večjo asimetrijo porazdelitve pojasnjevalnih spremenljivk dobili ožje intervale zaupanja in malenkost slabšo pokritost (razlika je približno 1 o.t. in je vidna, če primerjamo številske tabele).

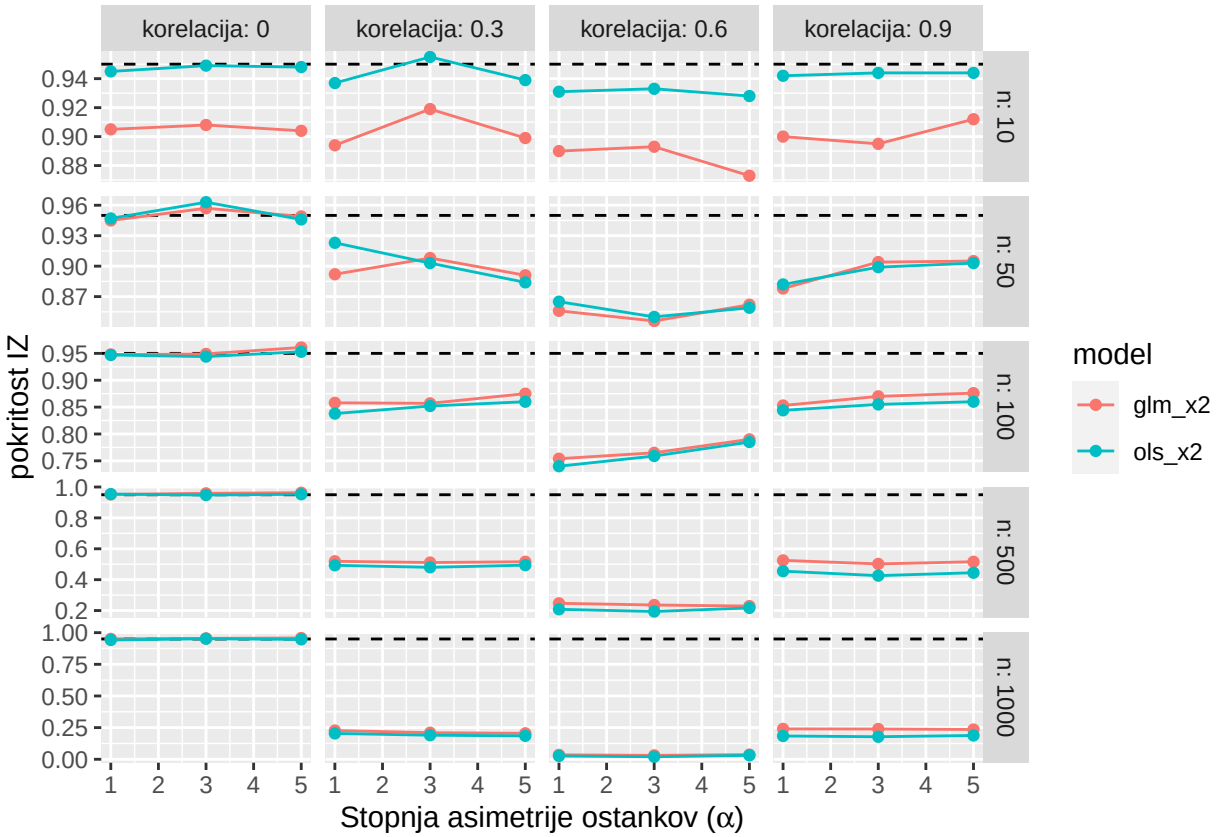
Tabela 11: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	glm_x2	glm_x3
10	0.3	0.91	0.90	0.91
10	0.9	0.90	0.90	0.91
100	0.3	0.95	0.86	0.95
100	0.9	0.96	0.87	0.96
1000	0.3	0.96	0.21	0.96
1000	0.9	0.95	0.24	0.96

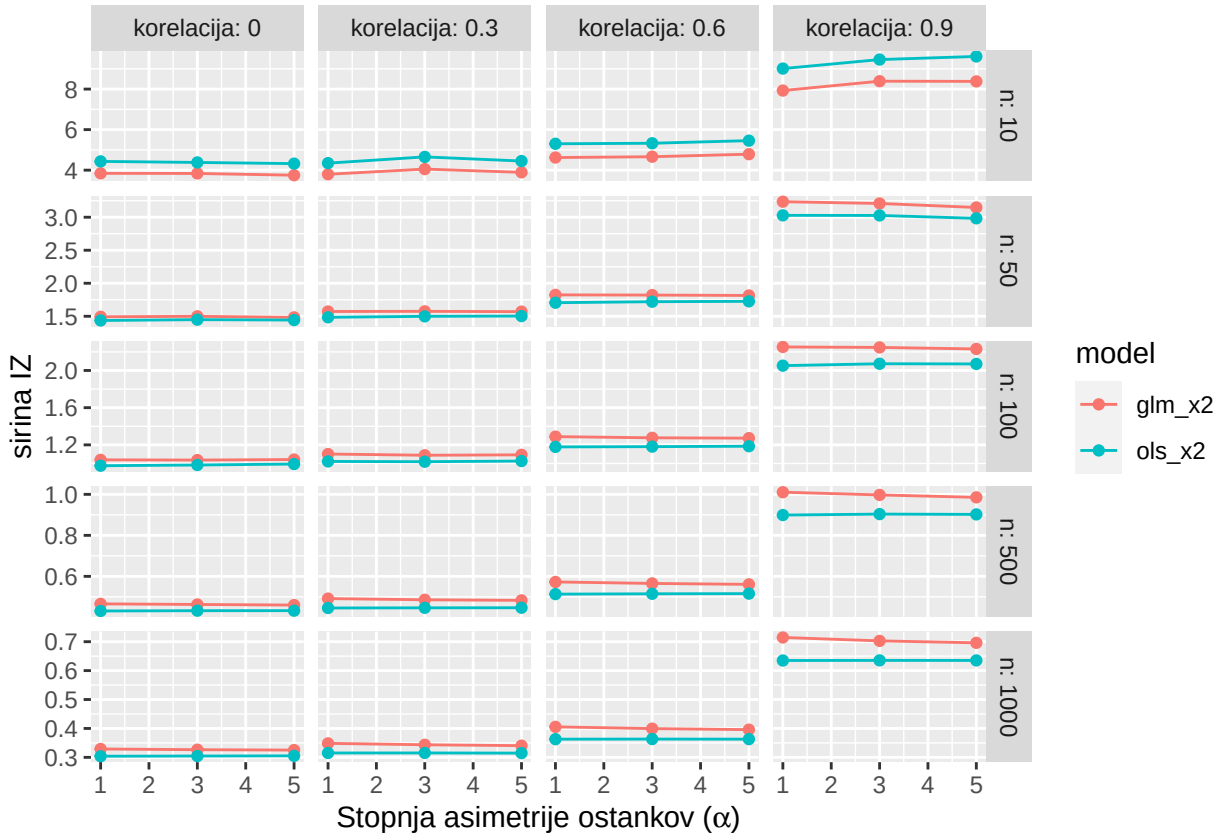
Tabela 12: Širina intervalov zaupanja za koeficient β_1 pri GLM in $Gamma(5, 5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm	glm_x2	glm_x3
10	0.3	3.96	3.91	3.60
10	0.9	9.97	8.23	8.08
100	0.3	1.03	1.09	1.01
100	0.9	2.51	2.24	2.21
1000	0.3	0.33	0.34	0.32
1000	0.9	0.79	0.70	0.69

Poglejmo si še razlike med metodama pri odstranjenih spremenljivkah. Najprej iz modela odstranimo X_2.



Slika 15: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(5, 5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 16: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

Pri pokritosti v modelih brez X_2 opazimo večje razlike v primerjavi z bolj asimetrično porazdelitvijo pojasnjevalnih spremenljivk. Tokrat imamo predvsem pri večjih vzorcih pokritost precej slabšo kot prej (primerjava tabel 7 in 13), vendar pa to lahko pripišemo predvsem ožjim intervalom zaupanja (primerjava tabel 8 in 14).

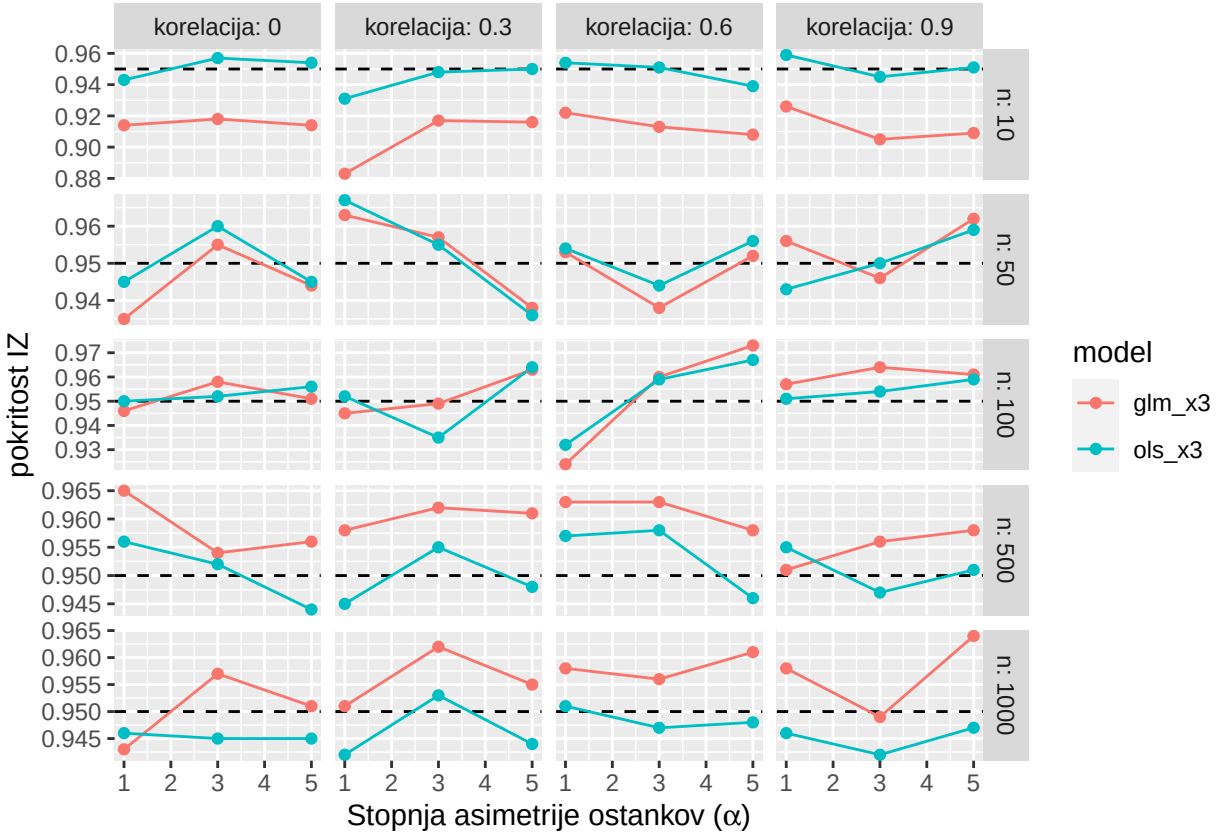
Tabela 13: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika
10	0.3	0.90	0.94	-0.04
10	0.9	0.90	0.94	-0.04
100	0.3	0.86	0.85	0.01
100	0.9	0.87	0.85	0.02
1000	0.3	0.21	0.19	0.02
1000	0.9	0.24	0.18	0.06

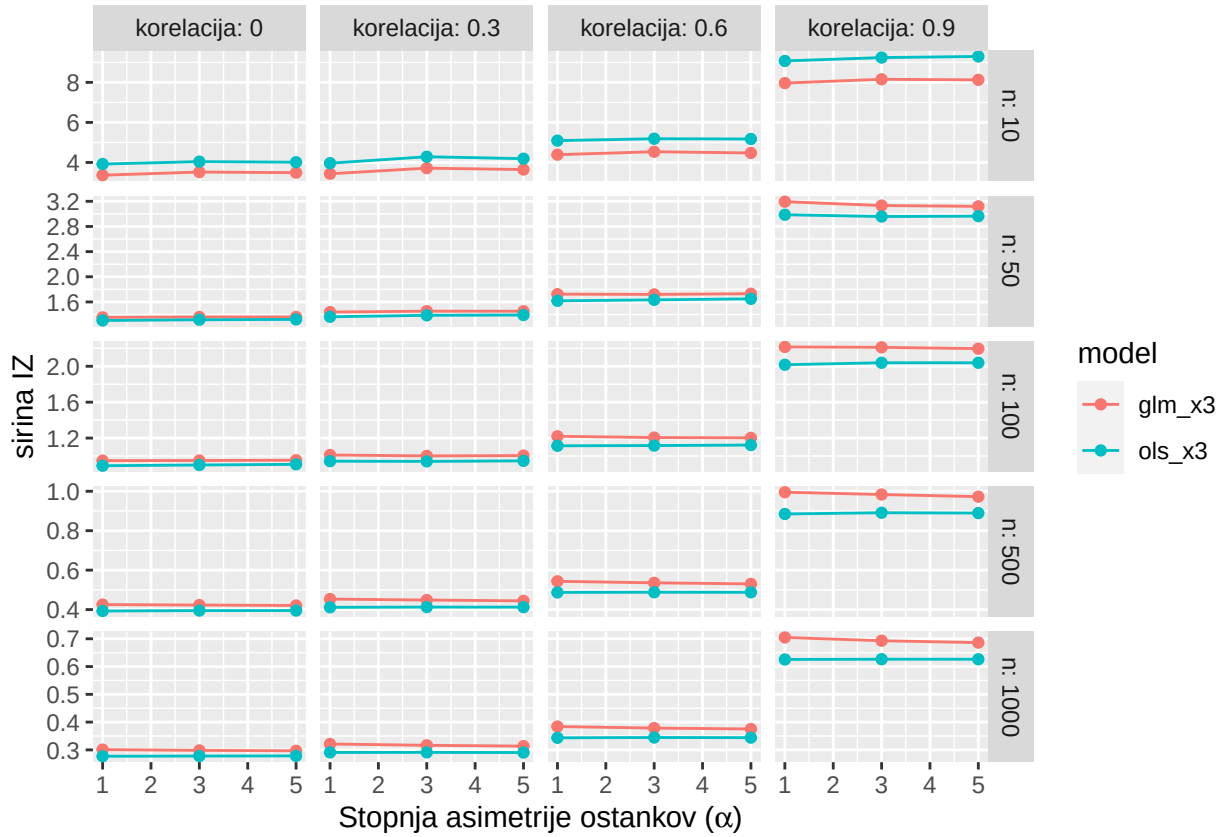
Tabela 14: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_2 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika	razlika_pct
10	0.3	3.91	4.48	-0.57	-0.13
10	0.9	8.23	9.36	-1.13	-0.12
100	0.3	1.09	1.02	0.07	0.07
100	0.9	2.24	2.06	0.18	0.09
1000	0.3	0.34	0.31	0.03	0.10
1000	0.9	0.70	0.64	0.06	0.09

Kaj pa se zgodi, če odstranimo X_3?



Slika 17: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk



Slika 18: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

Pri modelih brez spremenljivke X_3 v primeru bolj simetrične porazdelitve $Gamma(5,5)$ opazimo ožje intervale zaupanja kot pri $Gamma(2,5)$, pokritost ostaja zelo podobna (v nekaterih primerh je boljša, v drugih slabša za največ 1 o.t.).

Tabela 15: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x3	ols_x3	razlika
10	0.3	0.91	0.94	-0.03
10	0.9	0.91	0.95	-0.04
100	0.3	0.95	0.95	0.00
100	0.9	0.96	0.95	0.01
1000	0.3	0.96	0.95	0.01
1000	0.9	0.96	0.94	0.02

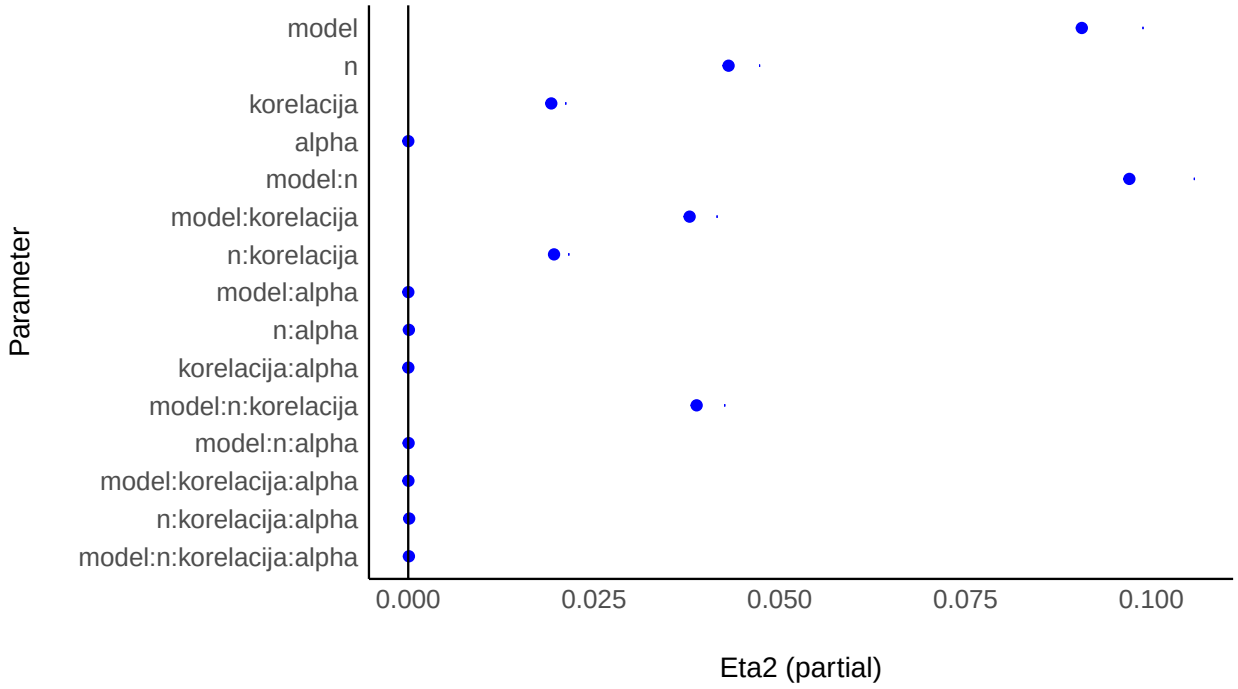
Tabela 16: Širina intervalov zaupanja za koeficient β_1 pri modelu brez X_3 in $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
10	0.3	3.60	4.15	-0.55	-0.13

n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
10	0.9	8.08	9.20	-1.12	-0.12
100	0.3	1.01	0.94	0.07	0.07
100	0.9	2.21	2.03	0.18	0.09
1000	0.3	0.32	0.29	0.03	0.10
1000	0.9	0.69	0.63	0.06	0.10

Analiza variance in velikost učinka

Ker smo že v prejšnjih poglavjih ugotovili, da pri pokritosti in širini intervalov zaupanja ni večjih razlik med posameznimi regresijskimi koeficienti, bomo analizo variance naredili le na rezultatih za koeficient β_1. Na spodnjih grafih je tako prikazana velikost učinka pri analizi varianci za pokritost in kasneje še za širino intervala zaupanja glede na vključene spremenljivke. Rezultate iz grafov beremo hierarhično, torej koliko variabilnosti dodatno pojasni posamezna spremenljivka, glede na že upoštevane spremenljivke.

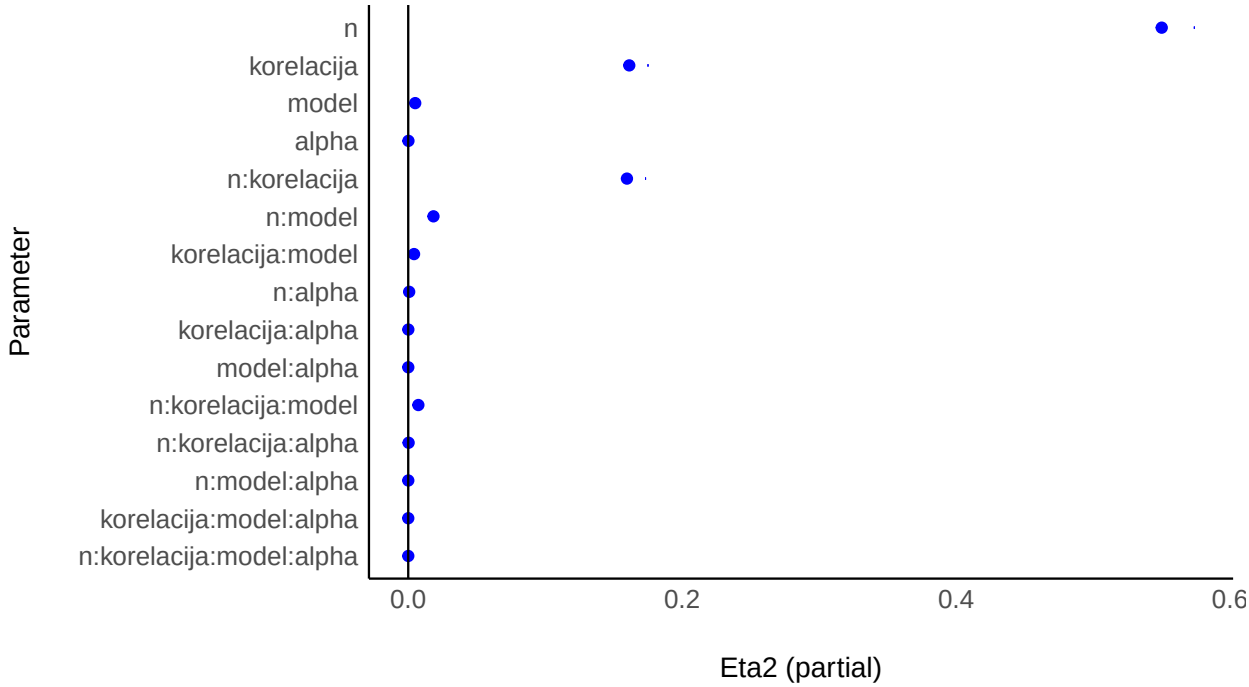


Slika 19: Velikost učinka pri analizi variance za pokritost intervala zaupanja

Iz zgornjega grafa lahko razberemo, da nobena od spremenljivk ne pojasni velikega deleža variabilnosti v pokritosti intervalov zaupanja. Vseeno pa je kar nekaj takih, ki pojasnijo manjši del in imajo statistično značilen vpliv (to nam pove povzetek modela anove, ki je v prilogi naloge). Na začetku model pojasni 9% variabilnosti pokritosti intervala zaupanja. Za tem velikost vzorca pojasni 4% preostale variabilnosti in korelacija še dodatna 2%. Asimetričnost ostankov za tem dodatno ne pojasni nič variabilnosti. Nekoliko večji del variabilnosti (glede na ostale vplive) dodatno pojasnijo še interakcija med modelom in velikostjo vzorca (10%), interakcija med modelom in korelacijo (4%), interakcija med korelacijo in velikostjo vzorca (2%) ter na koncu še interakcija modela, velikosti vzorca in korelacije (4%).

Pri tem se moramo zavedati, da znotraj spremenljivke `model` nista samo *glm* ali *lm*, ampak so upoštevani še nepolni modeli. To poudarimo zato, ker je učinek najverjetneje posledica nepolnih modelov. Tudi v prejšnjih prikazih smo zaznali vpliv sledečih modelov, medtem ko med *lm* in *glm* modeloma nismo opazili razlik. Vpliv variabilnosti in velikost vzorca smo prav tako zaznali že iz prejšnjih grafov, analiza variance in velikost učinka

pa nam naša opažanja še dodatno potrdita. Poglejmo si še velikost učinka pri analizi variance za širino intervalov zaupanja.



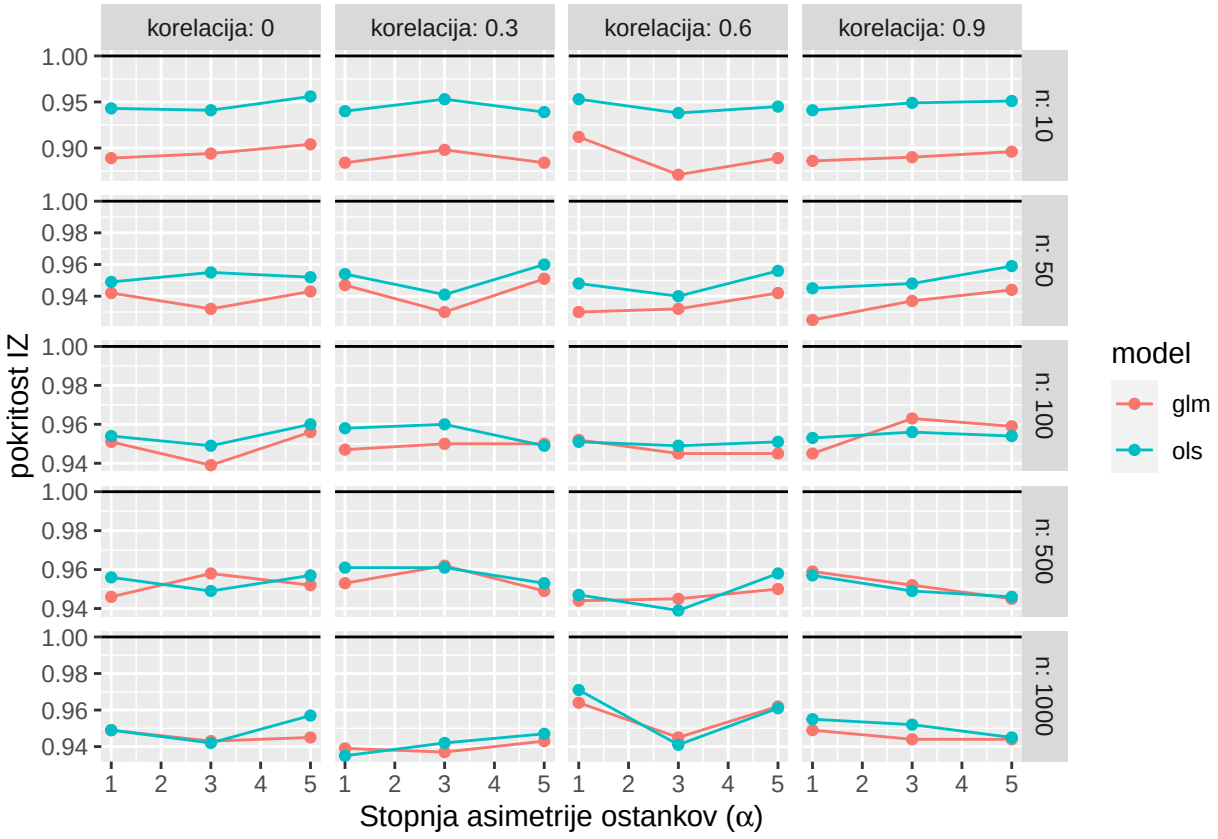
Slika 20: Velikost učinka pri analizi variance za širino intervala zaupanja

Opazimo, da ni toliko spremenljivk, ki bi pojasnjevale variabilnost širine intervalov zaupanja, vendar le te pojasnijo večji del variabilnosti. 55% variabilnost opazovane spremenljivke pojasni velikost vzorca, od preostale variabilnosti pa 16% pojasni korelacija med neodvisnimi spremenljivkami. Model in asimetričnost ostankov dodatno ne pojasnita večjega dela variabilnosti širine intervala zaupanja, dodatnih 16% pa pojasnjuje še interakcija med velikostjo vzorca in korelacijo. Manjši del variabilnosti dodatno pojasni še interakcija med velikostjo vzorca in modelom (2%), ostale interakcije pa zanemarljivo majhne deleže. Ponovno nam prikaz in analiza variance potrjujeta naša predhodna opazovanja.

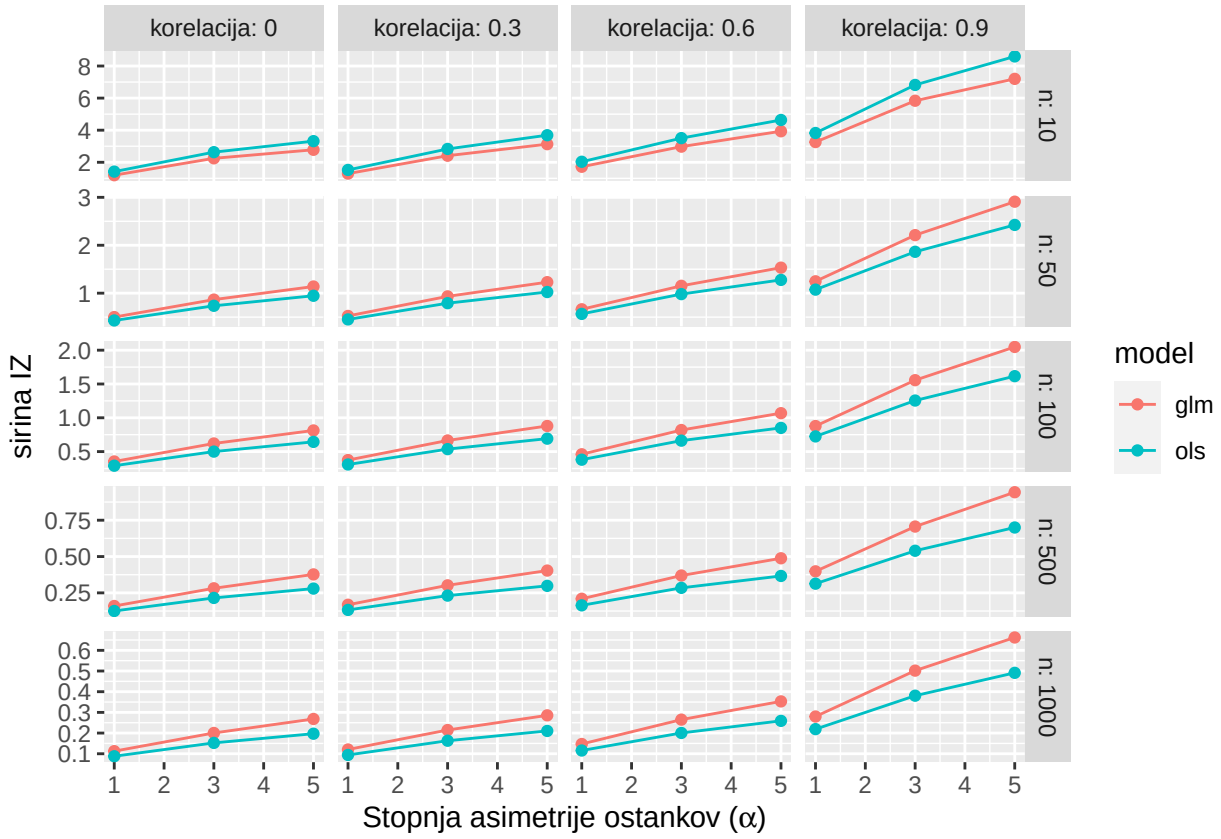
Transformacija odzivne spremenljivke

V tem poglavju so predstavljeni še rezultati simulacij, kjer je uporabljena logaritemska tranformacija. Pri teh simulacijah nas je zanimala predvsem razlika med uporabo transformacije $log(Y)$ v formuli funkcije lm ($log(Y) = \beta X$) in uporabo logaritemske link funkcije ($link =$ "log") v funkciji glm. Vseeno pa si bomo pogledali tudi razlike med rezultati pri različnih parametrih.

Zelo asimetrična porazdelitev pojasnjevalnih spremenljivk



Slika 21: Pokritost intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(2, 5)$



Slika 22: Širina intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(2, 5)$

Na Sliki 21 vidimo, da je pokritost IZ pri linearni regresiji boljša kot pri linearnih modelih. Razlika je očitna predvsem pri manjših vzorcih, pri dovolj velikih pa praktično ni razlik (številke so razvidne v tabeli 17). Širina intervalov zaupanja je pri obeh metodah precej podobna, vendar se linearna regresija vseeno izkaže za malenkost boljšo. Razlike se povečujejo predvsem z večanjem korelacije med pojasnjevalnimi spremenljivkami in z večanjem vzorca (kar je bolje vidno v tabeli 18). Predvidevamo, da je večanje širine intervalov zaupanja z večanjem stopnje asimetrije ostankov povezano s tem, da variance tokrat nismo skalirali. Vseeno pa lahko rečemo, da se *lm* model v primeru logaritemske transformacije pri $Gamma(2, 5)$ porazdelitvi pojasnjevalnih spremenljivk obnese bolje od *glm* modela tako pri pokritosti kot tudi širini IZ.

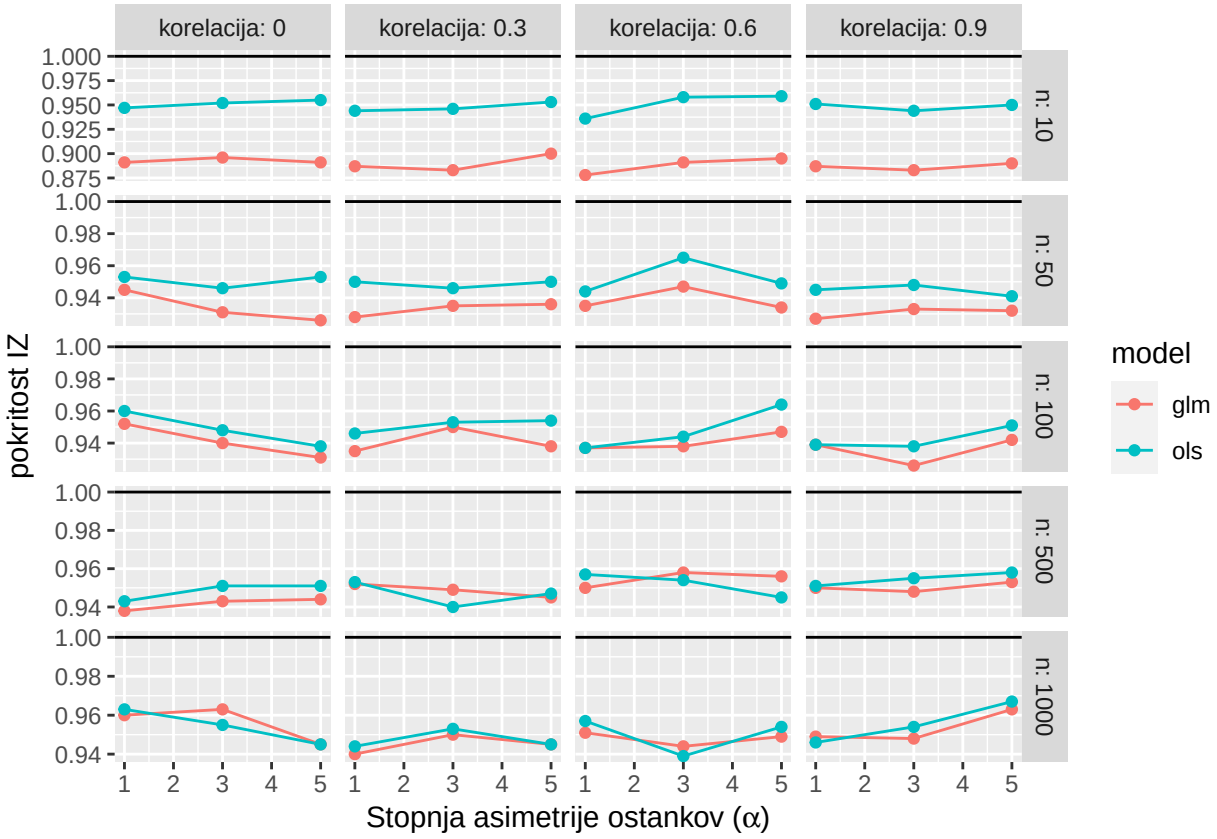
Tabela 17: Pokritost intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(2, 5)$

n	korelacija	glm	ols	razlika
10	0.3	0.89	0.94	-0.05
10	0.9	0.89	0.95	-0.06
100	0.3	0.95	0.96	-0.01
100	0.9	0.96	0.95	0.01
1000	0.3	0.94	0.94	0.00
1000	0.9	0.95	0.95	0.00

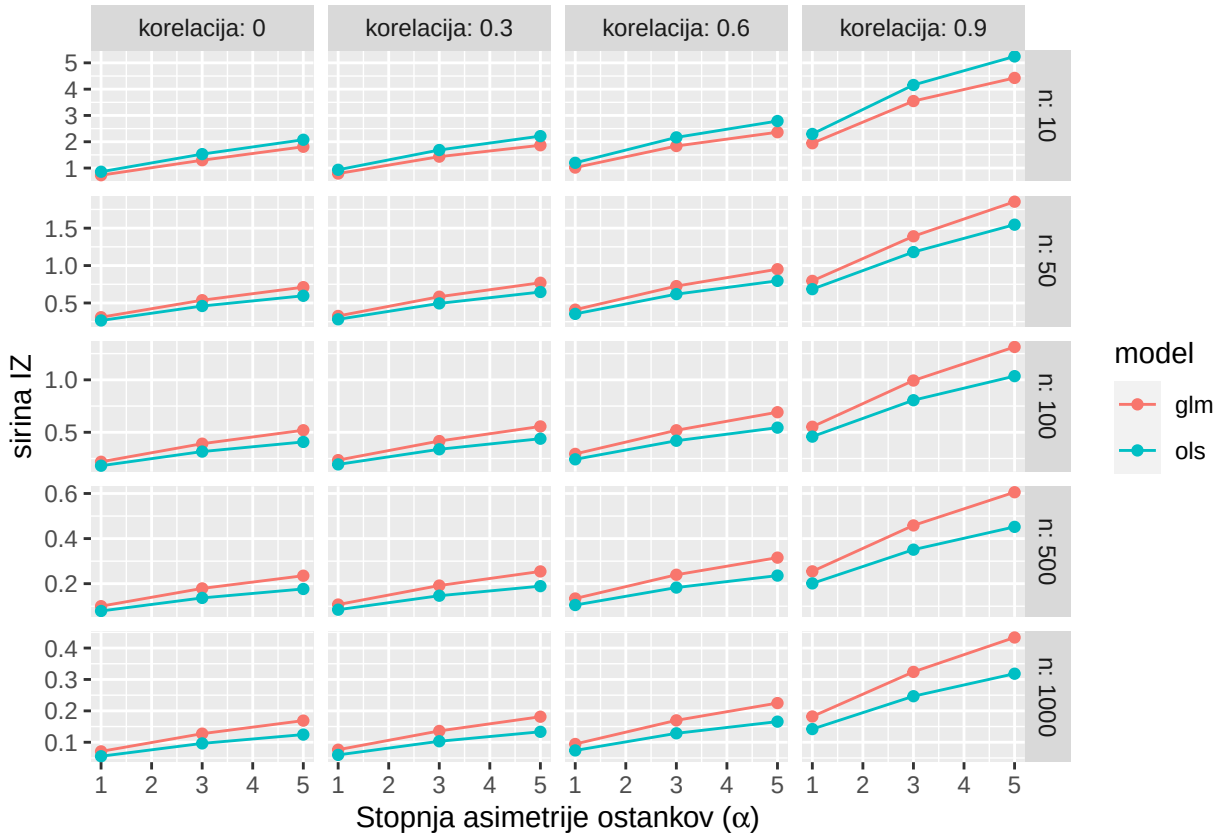
Tabela 18: Širina intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(2, 5)$

n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	2.28	2.68	-0.40	-0.15
10	0.9	5.43	6.41	-0.98	-0.15
100	0.3	0.64	0.51	0.13	0.25
100	0.9	1.49	1.20	0.29	0.24
1000	0.3	0.21	0.16	0.05	0.31
1000	0.9	0.48	0.36	0.12	0.33

Manj asimetrična porazdelitev pojasnjevalnih spremenljivk



Slika 23: Pokritost intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(5, 5)$



Slika 24: Širina intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(5,5)$

Pri $Gamma(5,5)$ porazdelitvi pojasnjevalnih spremenljivk pridemo do podobnih zaključkov kot pri $Gamma(2,5)$. Pokritost pri *lm* je boljša od pokritosti pri *glm* metodi, razlike so večje pri manjših vzorcih. Stopnja asimetrije ostankov tu ne igra vloge, prav tako korelacija med pojasnjevalnimi spremenljivkami. Slika 24 je zelo podobna Sliki 22, le da je skala širine IZ pomaknjena nekoliko nižje. Predvsem pri manjših vzorcih je širina IZ v primeru porazdelitve $Gamma(2,5)$ v povprečju očitno večja od širine v primeru $Gamma(5,5)$. To opazimo, ko primerjamo tabeli 18 in 20.

Tabela 19: Pokritost intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(5,5)$

n	korelacija	glm	ols	razlika
10	0.3	0.89	0.95	-0.06
10	0.9	0.89	0.95	-0.06
100	0.3	0.94	0.95	-0.01
100	0.9	0.94	0.94	0.00
1000	0.3	0.94	0.95	-0.01
1000	0.9	0.95	0.96	-0.01

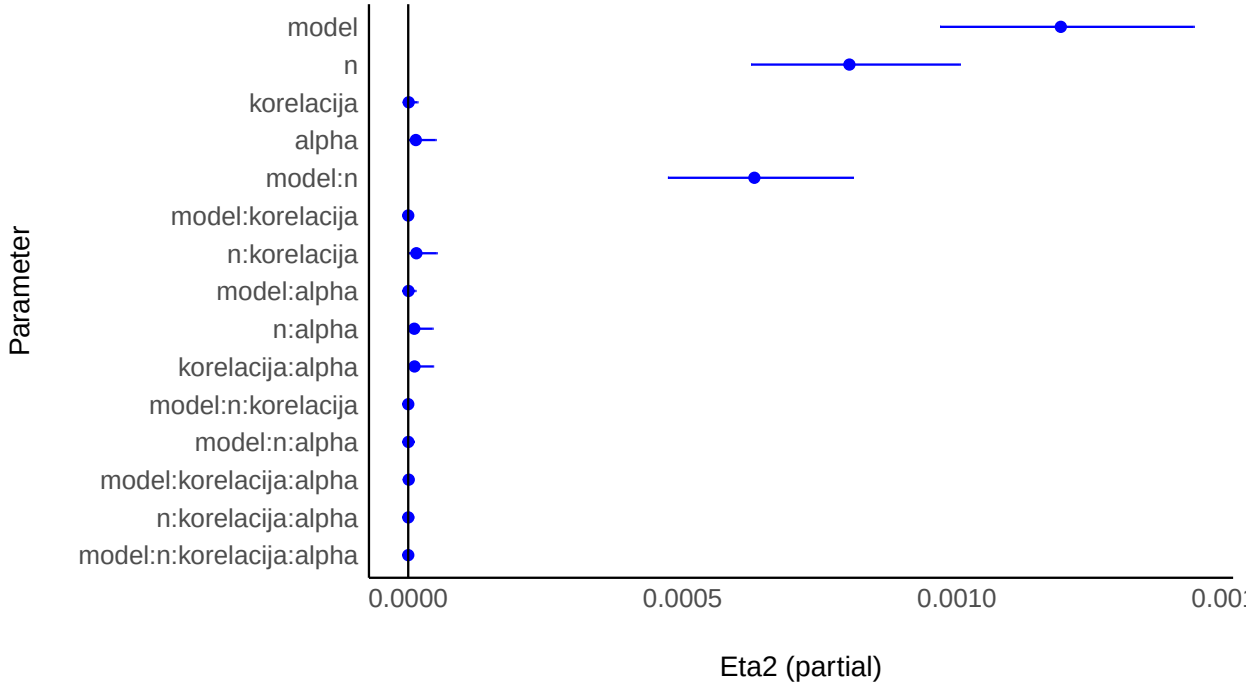
Tabela 20: Širina intervalov zaupanja za koeficient β_1 pri $X_i \sim Gamma(5,5)$

n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	1.36	1.61	-0.25	-0.16
10	0.9	3.31	3.90	-0.59	-0.15

n	korelacija	glm	ols	razlika	razlika_pct
100	0.3	0.40	0.32	0.08	0.25
100	0.9	0.95	0.77	0.18	0.23
1000	0.3	0.13	0.10	0.03	0.30
1000	0.9	0.31	0.24	0.07	0.29

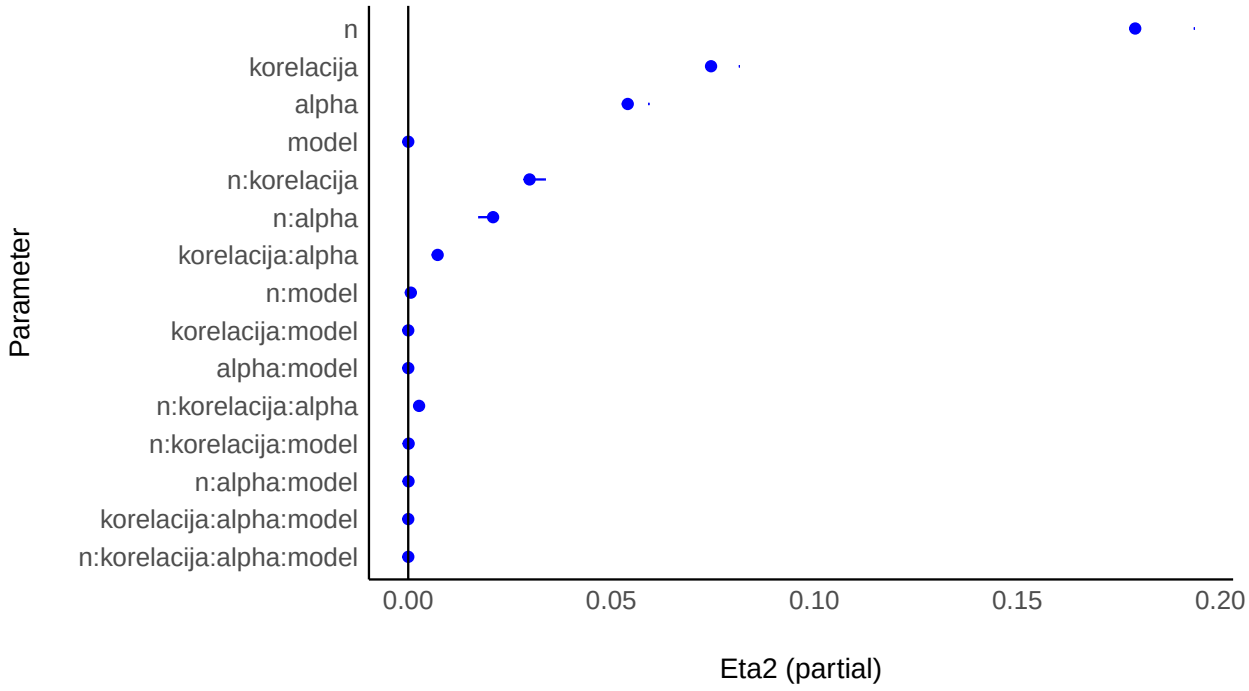
Analiza variance

Izvedemo podobno analizo kot pri modelih brez transformacije. Najprej si poglejmo analizo variance za pokritost, nato pa še za širino intervalov zaupanja.



Slika 25: Velikost učinka pri analizi variance za pokritost intervala zaupanja

Iz Slike 25 lahko razberemo, da nobena od spremenljivk ne pojasni večjega deleža variabilnosti pokritosti intervalov zaupanja. Vpliv modela, velikosti vzorca in interakcije med modelom in velikostjo vzorca imajo sicer statistično značilen vpliv (razvidno iz priloge), vendar vse spremenljivke skupaj pojasnijo manj kot 1% variabilnosti. Z danimi spremenljivkami pa lahko pojasnimo precej večji delež variabilnosti širine intervalov zaupanja, kar je razvidno na naslednji sliki.



Slika 26: Velikost učinka pri analizi variance za širino intervala zaupanja

Velikost vzorca pojasni 18% variabilnosti, korelacija dodatno 7% in zatem porazdelitev ostankov še dodatnih 5%. Pri porazdelitvi ostankov se moramo zavedati, da je to posledica različne variance, ne pa asimetrije porazdelitev. Model zatem dodatno pojasni zanemarljiv delež variabilnosti, interakcija med velikostjo vzorca in korelacijo dodatno 3% in korelacija med velikostjo vzorca še 2% variabilnosti širine intervalov zaupanja. Ostale interakcije dodatno pojasnijo zanemarljiv delež variabilnosti.

Ugotovitve

Največji vpliv na ocene regresijskih koeficientov imata korelacija med pojasnjevalnimi spremenljivkami in velikost vzorca. Z večanjem korelacije se večajo širine intervalov zaupanja, kar je pričakovano, saj lahko del vpliva ene spremenljivke prevzame druga spremenljivka. Z večanjem velikosti vzorca pri polnih modelih dobimo večjo pokritost pri *glm* modelu in ožje intervale zaupanja pri obeh modelih (*glm* in *lm*). Pri dovolj velikih vzorcih (tu izvzamemo $n = 10$) ima *glm* nekoliko boljšo pokritost, a širše intervale zaupanja, kot *lm* model. Pri najmanjšem opazovanem vzorcu ($n = 10$) se pristop z *lm* funkcijo izkaže za bolj učinkovitega pri pokritosti a slabšega pri širini intervalov zaupanja.

Pri večji asimetriji porazdelitve pojasnjevalnih spremenljivk dobimo nekoliko širše intervale zaupanja kot pri bolj simetrični porazdelitvi. Pri bolj simetrični porazdelitvi $Gamma(5, 5)$ so tudi razlike med pristopoma z *glm* in *lm* manjše.Ob vsakem primeru pa velja, da se (procentualna) razlika v širini z večanjem velikosti vzorca povečuje v prid modela *glm*.

Večja korelacija med pojasnjevalnimi spremenljivkami poveča širino intervalov zaupanja regresijskih koeficientov, na pokritosti pa nima večjega vpliva. Izjema so modeli brez spremenljivke X_2. Asimetrija porazdelitve ostankov na ocene regresijskih koeficientov nima posebnega vpliva. V primeru logaritemske transformacije smo opazili, da se širina IZ z večanjem α (v $Gamma(\alpha, 5)$ porazdelitvi ostankov) povečuje, vendar pa to lahko pripišemo le vplivu večje variance in ne asimetrije. Z večanjem velikosti vzorca se pričakovano ožajo intervali zaupanja regresijskih koeficientov. Na pokritost IZ velikost vzorca nima vpliva, razen v primeru modelov brez spremenljivke X_2.

Izločitev spremenljivke X_3 iz modela po pričakovanjih nima bistvenega vpliva na pokritost regresijskih koeficientov, saj smo podatke generirali pod predpostavko $\beta_3 = 0$. Vseeno pa v modelu brez X_3 dobimo nekoliko ožje intervale zaupanja, zaradi česar lahko tak model ocenimo kot najboljši (pravi). Rezultati so pričakovani, saj v modelih, kjer imamo vključene vse spremenljivke in neničelno korelacijo med pojasnjevalnimi spremenljivkami, del vpliva spremenljivke X_2 lahko prevzame spremenljivka X_3, zaradi česar prihaja do slabših rezultatov. Izločitev spremenljivke X_2 predvsem pri večjih vzorcih in neničelni korelaciji med pojasnjevalnimi spremenljivkami vpliva na slabšo pokritost IZ regresijskega koeficienta β_1. V primeru ničelne korelacije in zadostne velikosti vzorca izločitev katerekoli od spremenljivk (X_2 ali X_3) iz modela ne vpliva na pokritost IZ regresijskega koeficienta β_1.

Nobena od spremenljivk, ki smo jih uporabili v tej nalogi, ne pojasni velikega deleža variabilnosti pokritosti intervalov zaupanja. Vseeno pa je nekaj takih, ki pojasnijo manjši delež variabilnosti in imajo statistično značilen vpliv. To so velikost vzorca, model (predvsem izključevanje spremenljivk), korelacija in različne interakcije. Z manjšim številom spremenljivk pa lahko pojasnimo večji delež variabilnosti širine intervalov zaupanja. Kar 55% variabilnosti pojasni velikost vzorca, večji delež preostale variabilnosti pa lahko pojasnimo s korelacijo in interakcijo med velikostjo vzorca in korelacijo.

Pri modelih z logaritemsko transformacijo se je pristop s funkcijo *lm* izkazal za boljšega tako pri pokritosti kot širini intervalov zaupanja. Na tem področju bi bilo zanimivo preveriti še, kako se obnese *glm* funkcija v primeru enake formule kot v *lm* in zapisom $link =$ "*identity*" (torej $log(Y)$ namesto $link =$ "*log*"). Poleg tega bi bilo še smiselno preveriti, če bi se *link* funkcija morda bolje obnesla, če bi pri generiranju podatkov transformirali le ostanke, ne pa tudi pojasnjevalnih spremenljivk.

V modelih z logaritemsko transformacijo nobena od obravnavanih spremenljivk ne pojasni omembe vrednega deleža variabilnosti pokritosti IZ. Z določenimi spremenljivkami pa lahko pojasnimo večji delež variabilnosti širine IZ - te spremenljivke so velikost vzorca, korelacija, porazdelitev ostankov (varianca ostankov) in nekatere interakcije. Omeniti je še smiselno, da smo pri manj asimetrični porazdelitvi pojasnjevalnih spremenljivk $Gamma(5, 5)$ dobili ožje intervale zaupanja regresijskih koeficientov.

Če pogledamo vse ugotovitve skupaj, lahko sklenemo, da če nimamo opravka s transformacijami, če imamo dovolj velik vzorec in če imamo gama porazdelitev ostankov, ima *glm* model boljšo pokritost intervalov zaupanja, a na račun širših intervalov zaupanja. V primeru transformacije odzivne spremenljivke je bolj smiselno izbrati pristop, kjer že v vpisani formuli transformiramo odzivno spremenljivko, namesto da le to upoštevamo v *link* parametru funkcije *glm*. Seveda pa v praksi dobimo podatke, pri katerih sprva ne vemo, s kakšnimi transformacijami imamo opravka. V splošnem lahko rečemo, da je pristop z *lm* modelom povsem "konkurenčen" pristopu z *glm*, razlike v ocenah regresijskih koeficientov so relativno majhne. Delo P. E. Johnsona, ki je prišel do nekaterih podobnih ugotovitev, potrjuje naše sklepe. Sam je sklenil, da gama porazdelitev ostankov le redkokdaj vpliva na ocene regresijskih koeficientov, kljub temu, da so kršene nekatere predpostavke.

Viri

- P. E. Johnson, *GLM with a Gamma-distributed Dependent Variable*, [ogled 05.01.2020], dostopno na https://pj.freefaculty.org/guides/stat/Regression-GLM/Gamma/GammaGLM-01.pdf?fbclid=IwAR14W34VhGzyG0wPiqNTk1hWjIToAug6a2TsPsTeZKLj_ntfTxaR1Aowiko
- J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics, Springer Science + Business Media, LLC, New York, 2007.
- V. Maver, *Normalni linearni mešani modeli*, diplomsko delo, Fakulteta za matematiko in fiziko, Univerza v Ljubljani, 2018.
- *glm*, v: RDocumentation, [ogled 30.12.2020], dostopno na https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/glm.
- *lm*, v: RDocumentation, [ogled 30.12.2020], dostopno na https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm.

- *confint*, v: RDocumentation, [ogled 02.01.2021], dostopno na https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/confint
- M. Raič, *O linearni regresiji*, 2014. Najdeno na spletnem naslovu: http://valjhun.fmf.uni-lj.si/~raicm/Odlomki/Linearna_regresija.pdf
- L. Pfajfar, *Osnovna ekonometrija*, učbeniki Ekonomske fakultete, Ljubljana, 2018.
- *lcmix*, v: rdrr.io, [ogled 18.01.2021], dostopno na https://rdrr.io/rforge/lcmix/src/R/distributions.R
- *mvgamma*, v: rdrr.io, [ogled 18.01.2021], dostopno na https://rdrr.io/rforge/lcmix/man/mvgamma.html
- K. Siegrist, *5.8. The Gama Distribution*, [ogled 20.01.2021], dostopno na https://stats.libretexts.org/Bookshelves/Probability_Theory/Book%3A_Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/05%3A_Special_Distributions/5.08%3A_The_Gamma_Distribution

Priloge

Koda za simulacije brez transformacij se nahaja v priloženi datoteki `Simulacije.R`, njeni rezultati pa v datoteki `rezultati_simulacije.R`. Koda za simulacije s transformacijami se nahaja v priloženi datoteki `Simulacije_log.R`, njeni rezultati pa v datoteki `rezultati_simulacije_log.R`. Poročilo je bilo pripravljeno v `seminarska_RZM.Rmd` datoteki, ki vsebuje tudi kodo za grafe. Za pravilen prevod Rmd datoteke je potrebno v isti mapi imeti datoteko `header.tex`, ki poskrbi za lepši prevod datoteke (slovenski naslovi slik, tabel in podobno).

Sledi izpis ANOVE za pokritost intervalov zaupanja pri modelu brez transformacije.

```
##                              Df Sum Sq Mean Sq   F value   Pr(>F)
## model                         5   5139  1027.7 14346.406  < 2e-16 ***
## n                             4   2322   580.6  8104.944  < 2e-16 ***
## korelacija                    3   1012   337.2  4707.760  < 2e-16 ***
## alpha                         2      0     0.2     3.417   0.0328 *
## model:n                      20   5540   277.0  3867.053  < 2e-16 ***
## model:korelacija             15   2029   135.3  1888.147  < 2e-16 ***
## n:korelacija                 12   1032    86.0  1200.197  < 2e-16 ***
## model:alpha                  10      1     0.1     0.745   0.6821
## n:alpha                       8      4     0.6     7.705 2.22e-10 ***
## korelacija:alpha              6      1     0.1     1.659   0.1267
## model:n:korelacija           60   2082    34.7   484.285  < 2e-16 ***
## model:n:alpha                40      3     0.1     0.901   0.6487
## model:korelacija:alpha       30      1     0.0     0.522   0.9855
## n:korelacija:alpha           24      6     0.3     3.728 1.76e-09 ***
## model:n:korelacija:alpha    120      5     0.0     0.538   1.0000
## Residuals                719640  51551     0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sledi izpis ANOVE za širino intervalov zaupanja pri modelu brez transformacije.

```
##                              Df  Sum Sq Mean Sq   F value   Pr(>F)
## n                             4 4732191 1183048 2.200e+05  < 2e-16 ***
## korelacija                    3  744690  248230 4.616e+04  < 2e-16 ***
## model                         5   19781    3956 7.356e+02  < 2e-16 ***
## alpha                         2     532     266 4.950e+01  < 2e-16 ***
## n:korelacija                 12  735406   61284 1.140e+04  < 2e-16 ***
## n:model                      20   72611    3631 6.751e+02  < 2e-16 ***
## korelacija:model             15   16451    1097 2.039e+02  < 2e-16 ***
## n:alpha                       8    2605     326 6.054e+01  < 2e-16 ***
```

```
## korelacija:alpha              6     229      38 7.110e+00 1.36e-07 ***
## model:alpha                  10      88       9 1.632e+00   0.0907 .
## n:korelacija:model           60   28808     480 8.928e+01  < 2e-16 ***
## n:korelacija:alpha           24    1005      42 7.788e+00  < 2e-16 ***
## n:model:alpha                40      82       2 3.800e-01   0.9999
## korelacija:model:alpha       30      49       2 3.070e-01   0.9999
## n:korelacija:model:alpha    120     104       1 1.610e-01   1.0000
## Residuals                719640 3870180       5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sledi izpis ANOVE za pokritost intervalov zaupanja pri modelu s transformacijo.

```
##                             Df Sum Sq Mean Sq F value Pr(>F)
## model                        1     16  15.553 285.710 <2e-16 ***
## n                            1     11  10.510 193.077 <2e-16 ***
## korelacija                   1      0   0.010   0.192 0.6612
## alpha                        1      0   0.181   3.318 0.0685 .
## model:n                      1      8   8.245 151.458 <2e-16 ***
## model:korelacija             1      0   0.000   0.001 0.9750
## n:korelacija                 1      0   0.194   3.561 0.0591 .
## model:alpha                  1      0   0.005   0.090 0.7642
## n:alpha                      1      0   0.144   2.641 0.1041
## korelacija:alpha             1      0   0.150   2.749 0.0973 .
## model:n:korelacija           1      0   0.001   0.010 0.9207
## model:n:alpha                1      0   0.003   0.047 0.8276
## model:korelacija:alpha       1      0   0.012   0.212 0.6455
## n:korelacija:alpha           1      0   0.002   0.035 0.8522
## model:n:korelacija:alpha     1      0   0.000   0.004 0.9505
## Residuals               239984  13063   0.054
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sledi izpis ANOVE za širino intervalov zaupanja pri modelu s transformacijo.

```
##                             Df Sum Sq Mean Sq   F value   Pr(>F)
## n                            1  84670   84670 52328.667  < 2e-16 ***
## korelacija                   1  31288   31288 19336.907  < 2e-16 ***
## alpha                        1  22186   22186 13711.933  < 2e-16 ***
## model                        1      1       1     0.614    0.433
## n:korelacija                 1  11961   11961  7392.286  < 2e-16 ***
## n:alpha                      1   8286    8286  5120.740  < 2e-16 ***
## korelacija:alpha             1   2831    2831  1749.371  < 2e-16 ***
## n:model                      1    248     248   153.150  < 2e-16 ***
## korelacija:model             1      1       1     0.333    0.564
## alpha:model                  1      1       1     0.530    0.467
## n:korelacija:alpha           1   1039    1039   642.159  < 2e-16 ***
## n:korelacija:model           1     36      36    22.435 2.17e-06 ***
## n:alpha:model                1     25      25    15.490 8.29e-05 ***
## korelacija:alpha:model       1      0       0     0.003    0.955
## n:korelacija:alpha:model     1      4       4     2.534    0.111
## Residuals               239984 388302       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```