

Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2021-01-11

Kazalo vsebine

Uvod	2
Teoretični del	2
Posplošeni linearni modeli	2
Linearna regresija	2
Metoda najmanjših kvadratov (MNK)	3
Metoda iterativnega uteženega povprečja najmanjših kvadratov (IWLS)	3
Generiranje podatkov	4
Parametri	4
Funkciji lm() in glm()	5
Ocenjevanje intervalov zaupanja	5
Pričakovanja	5
Predstavitev rezultatov	6
Polni model	7
Vpliv odstranjevanja spremenljivk	12
Analiza variance in velikost učinka	24
Ugotovitve	25
Viri	26
Priloge	26

Uvod

Obravnavana metoda je linearna regresija in zanima nas, kako kršenje predpostavk (konkretnije 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.

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:

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

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 β_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 : $Cov(x_2, u) = Cov(x_3, u) = \dots = Cov(x_k, u) = 0$
- 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: $\lambda_1 X_1 + \lambda_2 X_2 + \dots + \lambda_k X_k = 0$
- 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 $N(\beta_1 x_{1i} + \dots + \beta_k x_{ki}, \sigma_u^2)$

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

$$(Y - X\beta)'W(Y - X\beta), \quad (1)$$

kjer je W znana simetrična matrika uteži.

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

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

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

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

Izkaže se, da je v smislu čim manjše variance optimalna izbira za matriko W matrika $W = V^{-1}$, kjer je $V = \text{Var}(Y)$. 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):

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

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 $Y_{ij}, j = 1, \dots, \tilde{m}$, 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 $t_1, \dots, t_{\tilde{m}}$. Za i -tega posameznika se lahko vektor meritev zapiše kot $Y_i = (Y_{ij})_{j \leq \tilde{m}}, i = 1, \dots, n$. Naj bodo Y_1, \dots, Y_n med seboj neodvisni in naj za njih velja $E(Y_i) = X_i\beta$ in $\text{Var}(Y_i) = V_0$. 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 $V = \text{diag}(V_0, \dots, V_0)$. 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

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

kjer je

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

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:

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

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: $X_j \sim \text{Gamma}(\delta, 5)$, $j = 1, 2, 3$, $\delta = 2, 5$;
- porazdelitev napak ($\text{Gamma}(\alpha, 5)$), 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 $X_i \sim \text{Gamma}(k_i, \theta)$, potem je

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

Ker zaradi spreminjanja parametrov v gama porazdelitvi pri porazdelitvi napak ne spreminjamo samo asimetričnosti ampak tudi varianco in ker vemo, da ima različna varianca 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.

Pri pregledu literature sva ugotovili, da za generiranje odvisnih gama spremenljivk lahko uporabimo kar funkcijo `rmvgamma()` 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. **Manjka še opis, kako funkcija generira vrednosti.**

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

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

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. Funkcija parametre modela ocenjuje po metodi iterativnega uteženega povprečja najmanjših kvadratov (`glm` iz RDocumentation, 2020).

Ocenjevanje intervalov zaupanja

Za ocenjevanje intervalov zaupanja pri obeh metodah uporabimo funkcijo `confint.default`, ki računa intervale zaupanja na podlagi standardnih napak. Najprej smo poskusili z uporabo navadne funkcije `confint`, vendar je prihajalo do problemov pri posplošenih linearnih modelih. Slednja funkcija predpostavlja normalnost, funkcija `confint.default` pa temelji na asimptotski normalnosti (`confint` iz RDocumentation, 2020). Za stopnjo zaupanja vzamemo $\alpha = 0.05$.

Pričakovanja

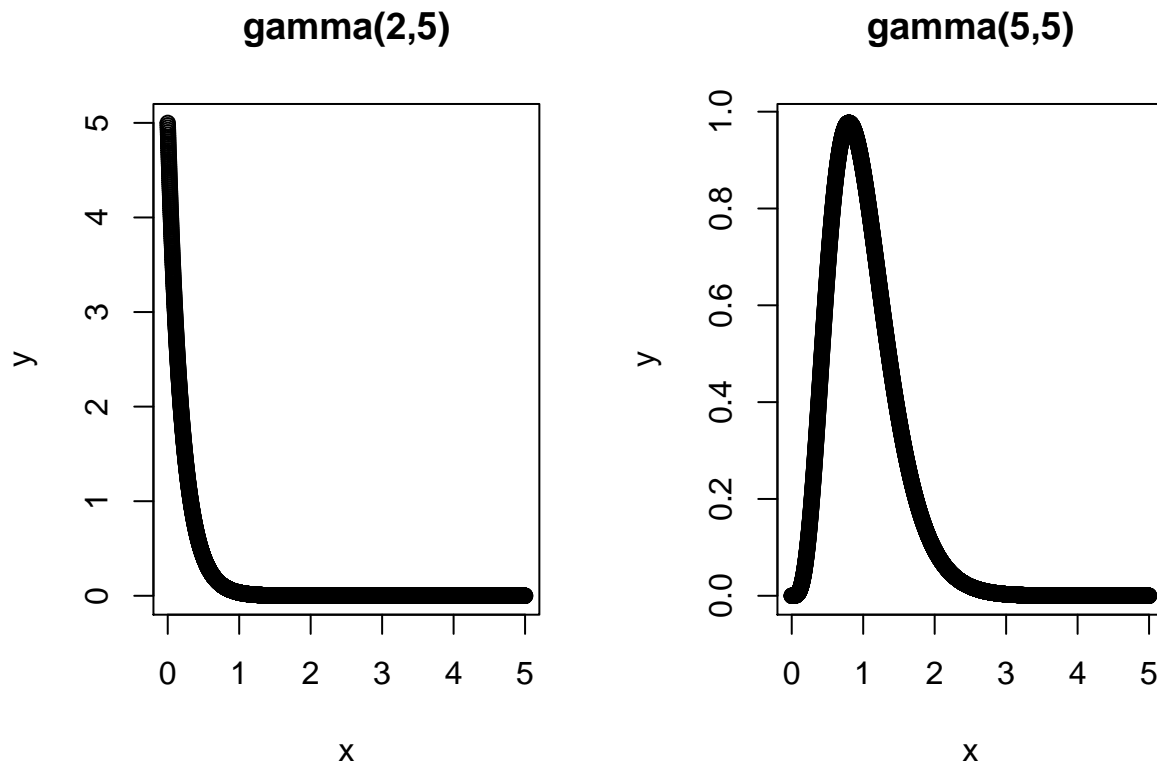
Pri večji korelaciji med pojasnjevalnimi spremenljivkami pričakujemo širše intervale zaupanja regresijskih koeficientov ne glede na izbiro metode. Zaradi večje širine intervalov zaupanja ne pričakujemo večjih sprememb v pokritosti.

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 , ne pričakujemo posebnih sprememb v rezultatih, saj spremenljivka nima vpliva na vrednost pojasnjevalne spremenljivke. V primeru, ko izločimo spremenljivko X_2 , pa pričakujemo spremembe v rezultatih - širše intervale zaupanja regresijskih koeficientov in slabšo pokritost.

Porazdelitev pojasnjevalnih spremenljivk preverimo za dve porazdelitvi - $\text{gama}(2, 5)$, ki je precej asimetrična in $\text{gama}(5, 5)$, 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.



Slika 1: Različni porazdelitvi gama

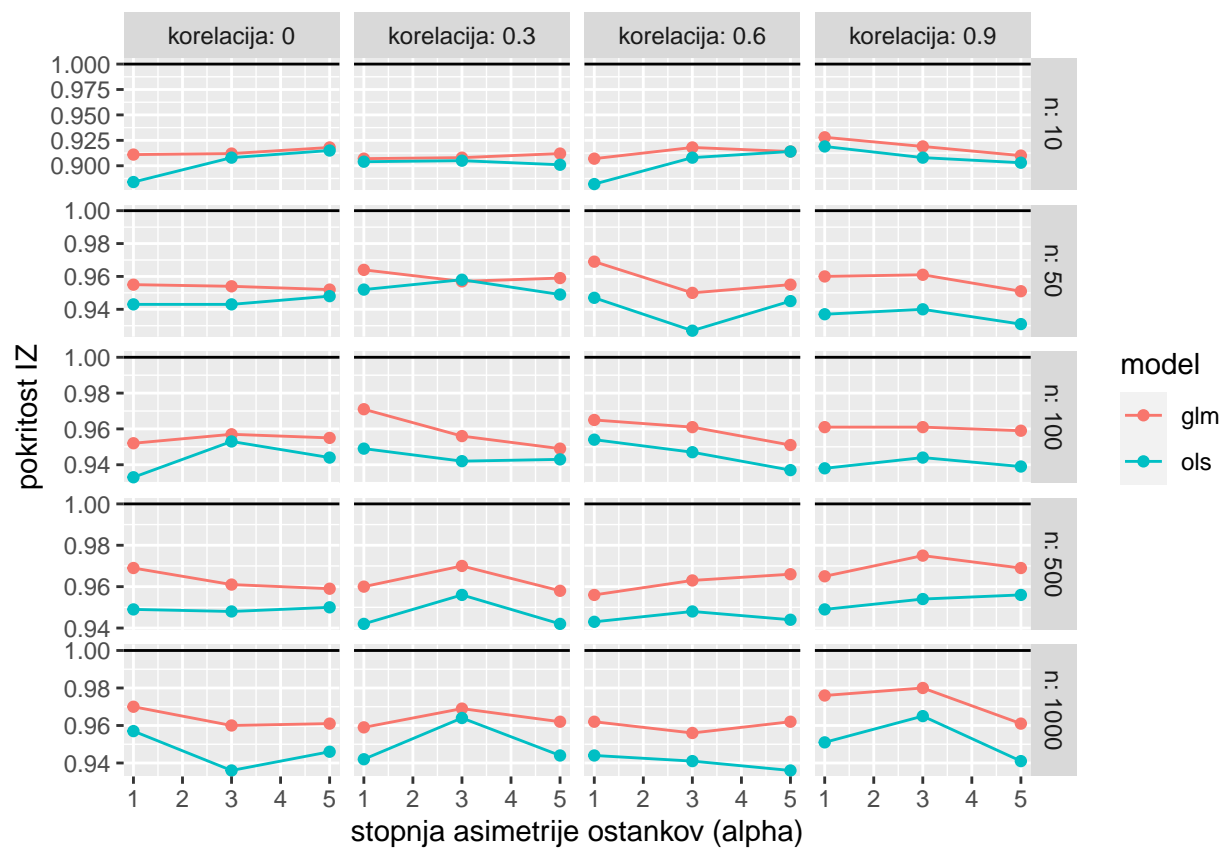
Predstavitev rezultatov

Rezultati se nanašajo na veliko parametrov in za več različnih načinov modeliranja, zato si bomo rezultate 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 .

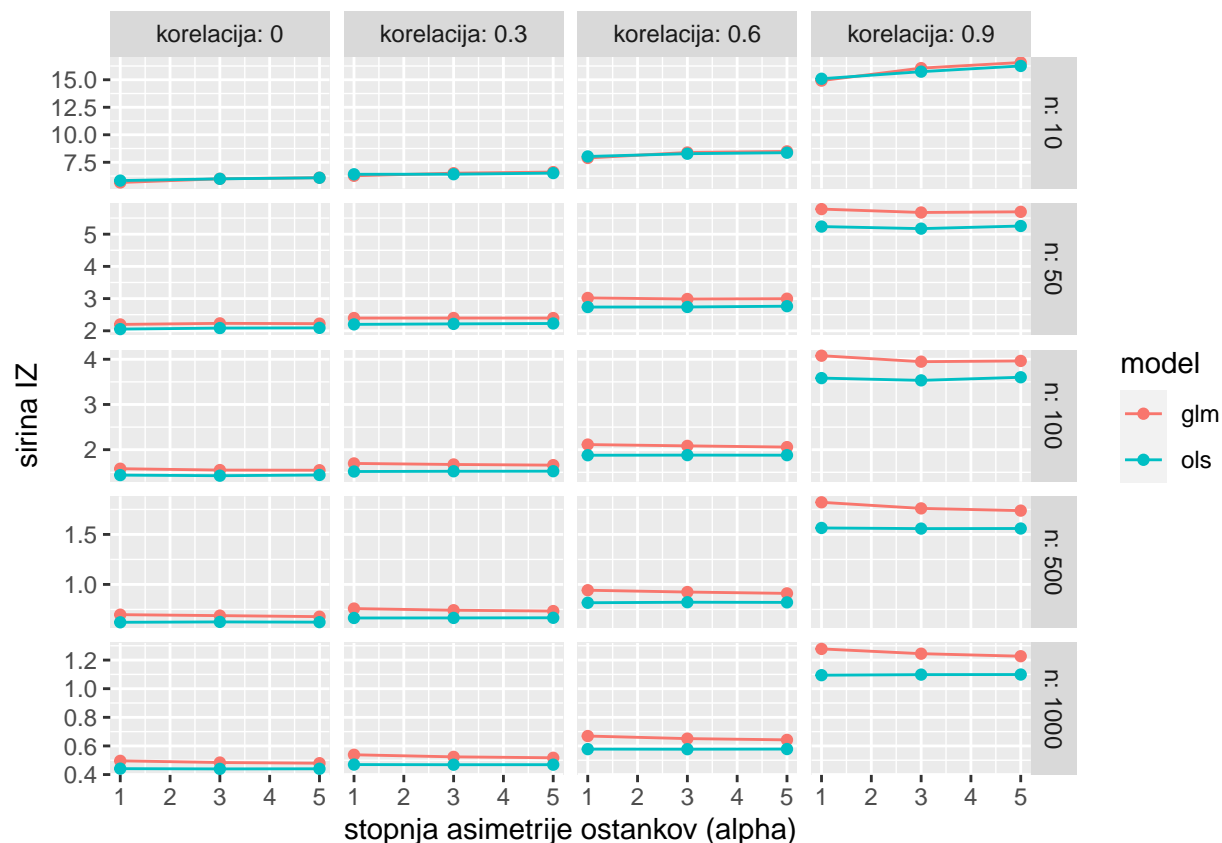
Polni model

Asimetrična porazdelitev pojasnjevalnih spremenljivk

Za začetek si pogledjmo rezultate oz. izbrane mere na polnem modelu, kjer so upoštevane vse spremenljivke in pri asimetrični porazdelitvi spremenljivke X . Naslednje dva grafa tako prikazujeta pokritost in širine intervalov zaupanja za koeficient β_1 za posamezen model (`glm` ali `ols`) glede na velikost vzorca in korelacijo.



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



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

Iz slik lahko vidimo, da imata precej očiten vpliv pri obeh metodah (glm in ols) korelacija in velikost vzorca. Z večanjem korelacije se sama pokritost ne spremeni bistveno, se pa močno povečajo intervali zaupanja, medtem ko z večanjem velikosti vzorca pridobimo večjo pokritost in ožje intervale zaupanja. Pri stopnji asimetrije ostankov pa ni opaziti bistvenega vpliva na rezultate.

Razberemo tudi, da ima model glm nekoliko boljšo pokritost, razlika pa je manjša pri manjših vzorcih. 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 torej preozki, posledično pa interval zaupanja za koeficient večkrat ne vsebuje prave vrednosti koeficienta. Razlike v pokritosti sicer niso velike, vseeno pa so prisotne. Rezultati so za manjše število izbranih parametrov n in korelacije prikazani tudi v spodnjih dveh tabelah, kjer bomo lažje razbrali razlike. Prikazane razlike so v obeh primerih razlike glm glede na ols metodo, v tabeli s širino intervalov zaupanja so prikazane razlike še v deležu glede na širino intervala zaupanja pri metodi ols.

n	korelacija	glm	ols	razlika
10	0.3	0.91	0.90	0.01
10	0.9	0.92	0.91	0.01
100	0.3	0.96	0.94	0.02
100	0.9	0.96	0.94	0.02
1000	0.3	0.96	0.95	0.01
1000	0.9	0.97	0.95	0.02

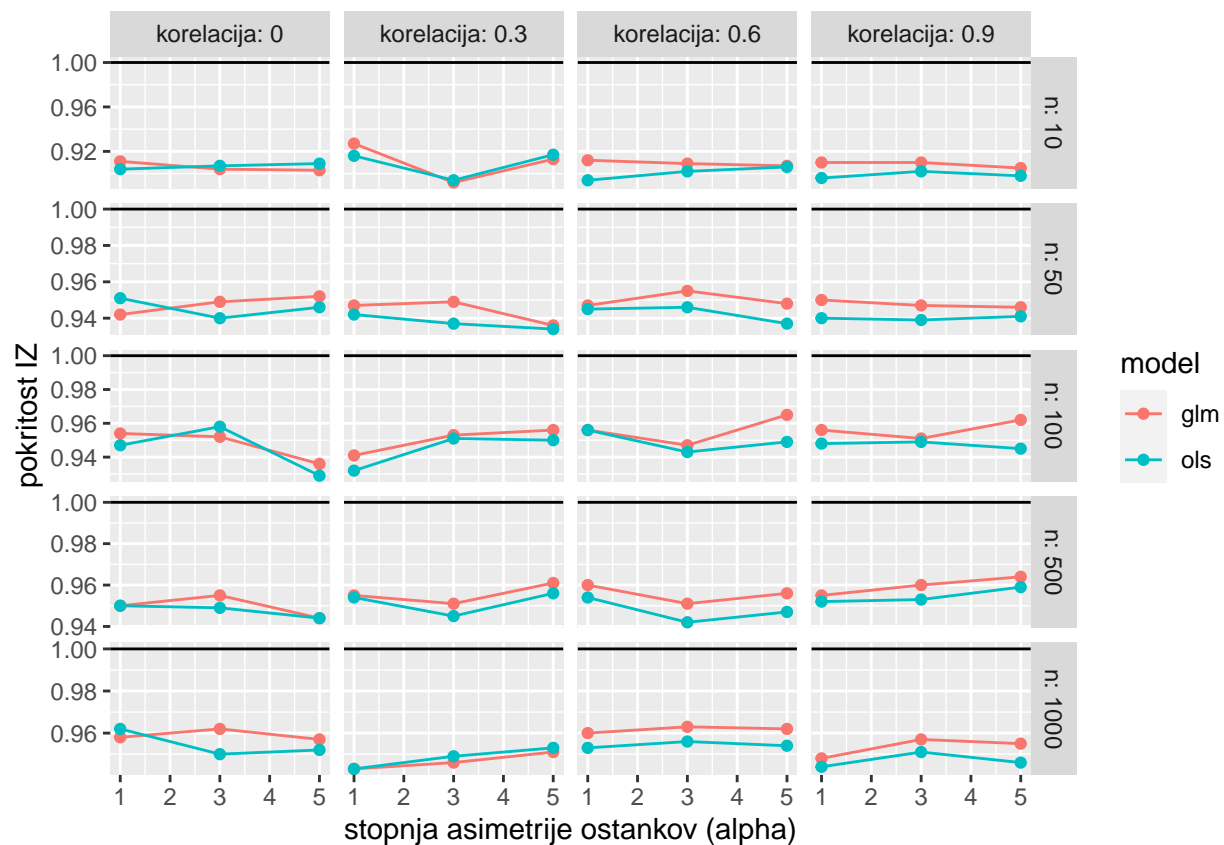
n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	6.47	6.45	0.02	0.00
10	0.9	15.85	15.69	0.16	0.01
100	0.3	1.68	1.52	0.16	0.11
100	0.9	4.00	3.57	0.43	0.12
1000	0.3	0.53	0.47	0.06	0.13
1000	0.9	1.25	1.10	0.15	0.14

Velja torej, da ima v povprečju glm metoda od 1 do 2 o.t. boljšo pokritost, vendar pa ima tudi do 14% širše intervale zaupanja. Ta razlika med intervali zaupanja se najbolj poveča z večanjem velikosti vzorca. V splošnem pa velja, da pokritost z večanjem velikosti vzorca zraste za 4%. Visoka korelacija v kombinaciji z majhnim vzorcem je najbolj problematična, tam imamo tudi manjšo pokritost in precej široke intervale zaupanja.

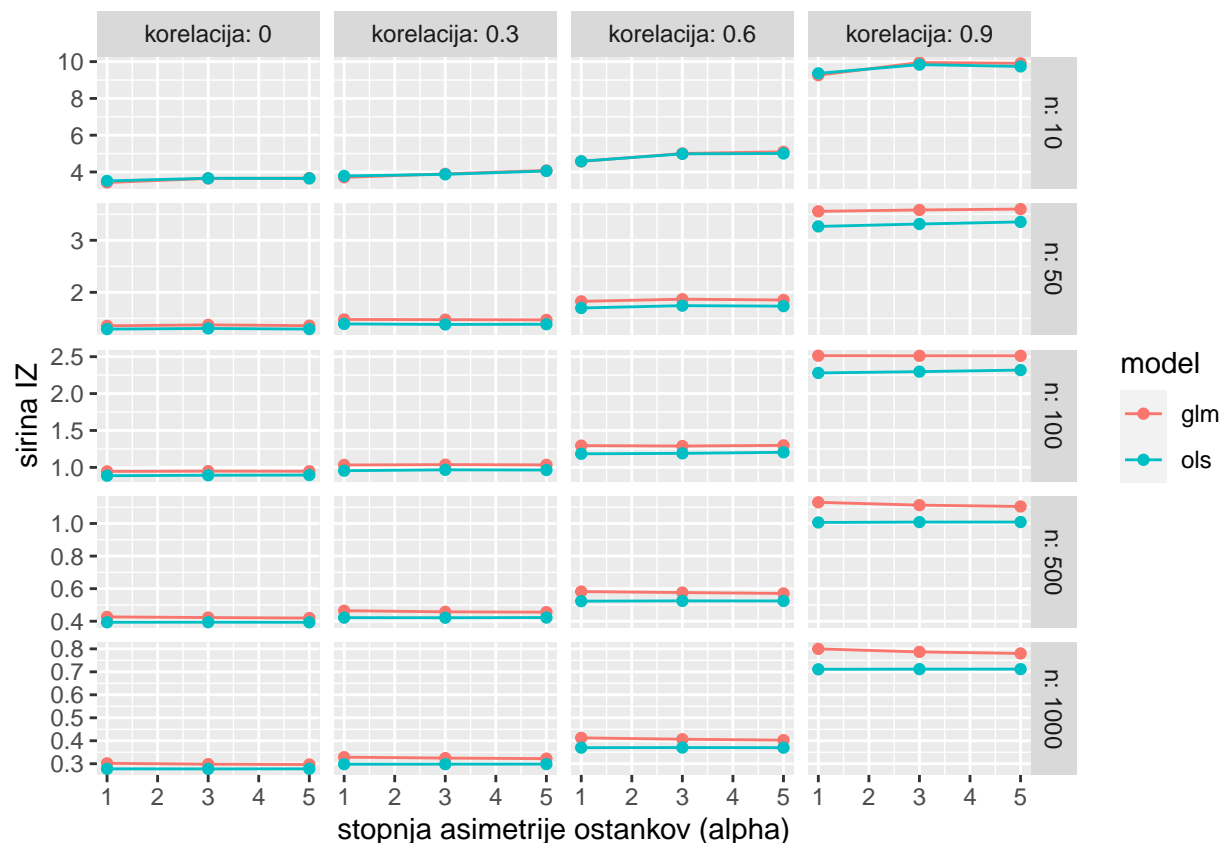
Prejšnji komentar: Na Sliki 1 vidimo, da so pokritosti intervalov zaupanja za regresijske koeficiente β_1 , β_2 in β_3 vedno blizu 100%, ne glede na uporabljeno metodo, velikost vzorca, korelacijo med pojasnjevalnimi spremenljivkami in stopnjo asimetrije ostankov. Močno izstopa pokritost konstante. Pri večjem vzorcu je njena pokritost praktično 0, pri manjših vzorcih pa se opazi vpliv asimetrije ostankov. Bolj ko je porazdelitev ostankov simetrična ($gamma(5, 5)$), slabša je pokritost koeficienta β_0 . Predvidevamo, da to niti ni posledica (a)simetričnosti, pač pa pričakovane vrednosti ostankov. Večja kot je pričakovana vrednost ostankov, slabša je pokritost intervalov zaupanja koeficienta β_0 . Pri velikosti vzorca $n = 10$ lahko vidimo, da pokritost pada praktično linearno z večanjem parametra α v porazdelitvi ostankov. Velikost korelacije med pojasnjevalnimi spremenljivkami prav tako vpliva na pokritost koeficienta β_0 - večja kot je korelacija, slabša je pokritost.

Manj simetrična porazdelitev pojasnjevalnih spremenljivk

Enako kot smo si pogledali za asimetrično porazdelitev pojasnjevalnih spremenljivk si pogledajmo še za nekoliko bolj simetrično porazdelitev.



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



Slika 5: 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 da sama asimetrija ostankov nima bistvenega vpliva. Razlike med metodama sta tokrat nekoliko manj opazni, sploh kar se tiče pokritosti, za razlike v širinah intervalov zaupanja pa si moramo pogledati številke v tabelah, kjer bomo lažje razbrali ali so le-te manjše ali večje kot prej. Predvsem lahko opazimo, če primerjamo graf širine intervalov zaupanja s prejšnjih grafom kjer je upoštevana asimetrična porazdelitev, da so intervali v splošnem nekoliko ožji.

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

n	korelacija	glm	ols	razlika	razlika_pct
10	0.3	3.90	3.90	0.00	0.00
10	0.9	9.71	9.65	0.06	0.01
100	0.3	1.03	0.96	0.07	0.07
100	0.9	2.51	2.30	0.21	0.09
1000	0.3	0.33	0.30	0.03	0.10

n	korelacija	glm	ols	razlika	razlika_pct
1000	0.9	0.79	0.71	0.08	0.11

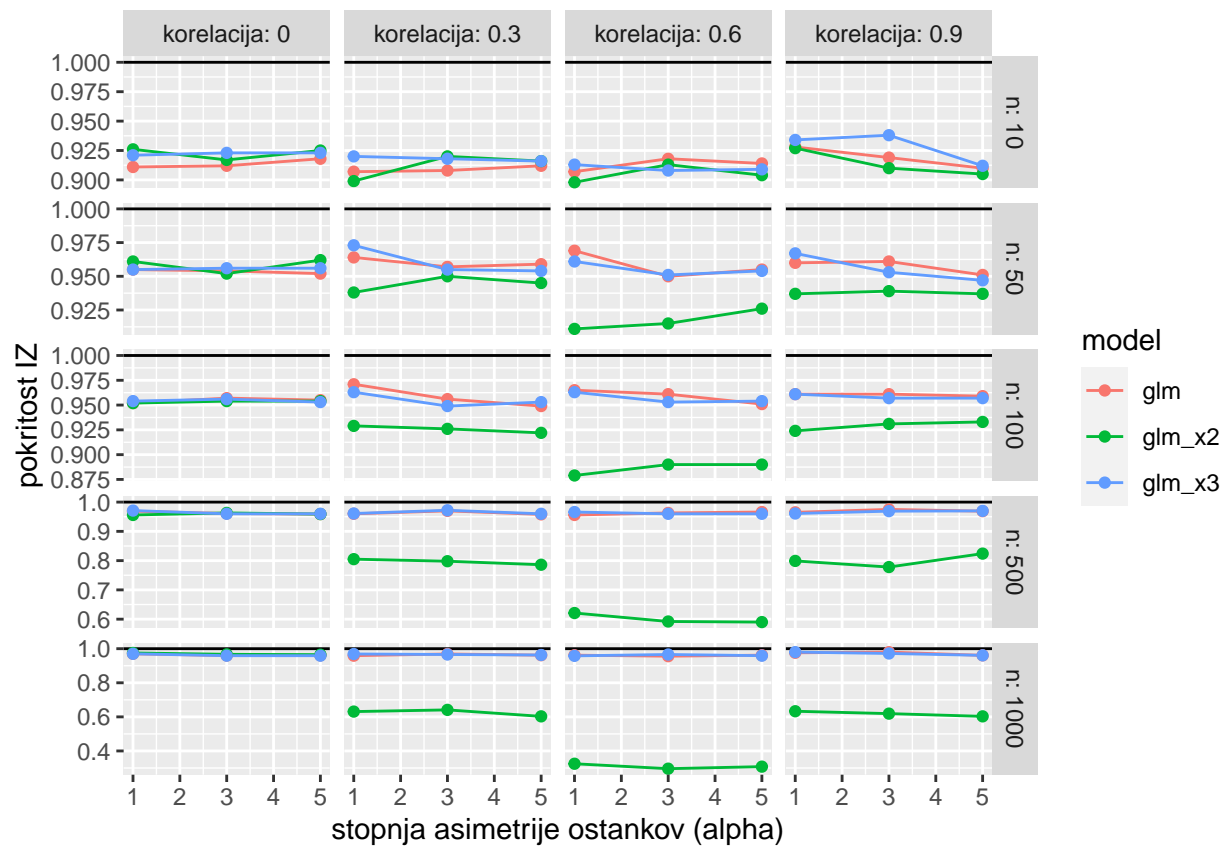
Številke v tabelah potrjujejo naša opažanja. Pri pokritosti je med metodama `glm` in `ols` manjša razlika in sicer je tokrat pokritost pri metodi `glm` višja za 0 do 1 o.t. Prav tako je manjša razlika v širini intervalov zaupanja. Ponovno pa velja, da se razlika v širini (procentualna) viša z večanjem velikosti vzorca.

Vpliv odstranjevanja spremenljivk

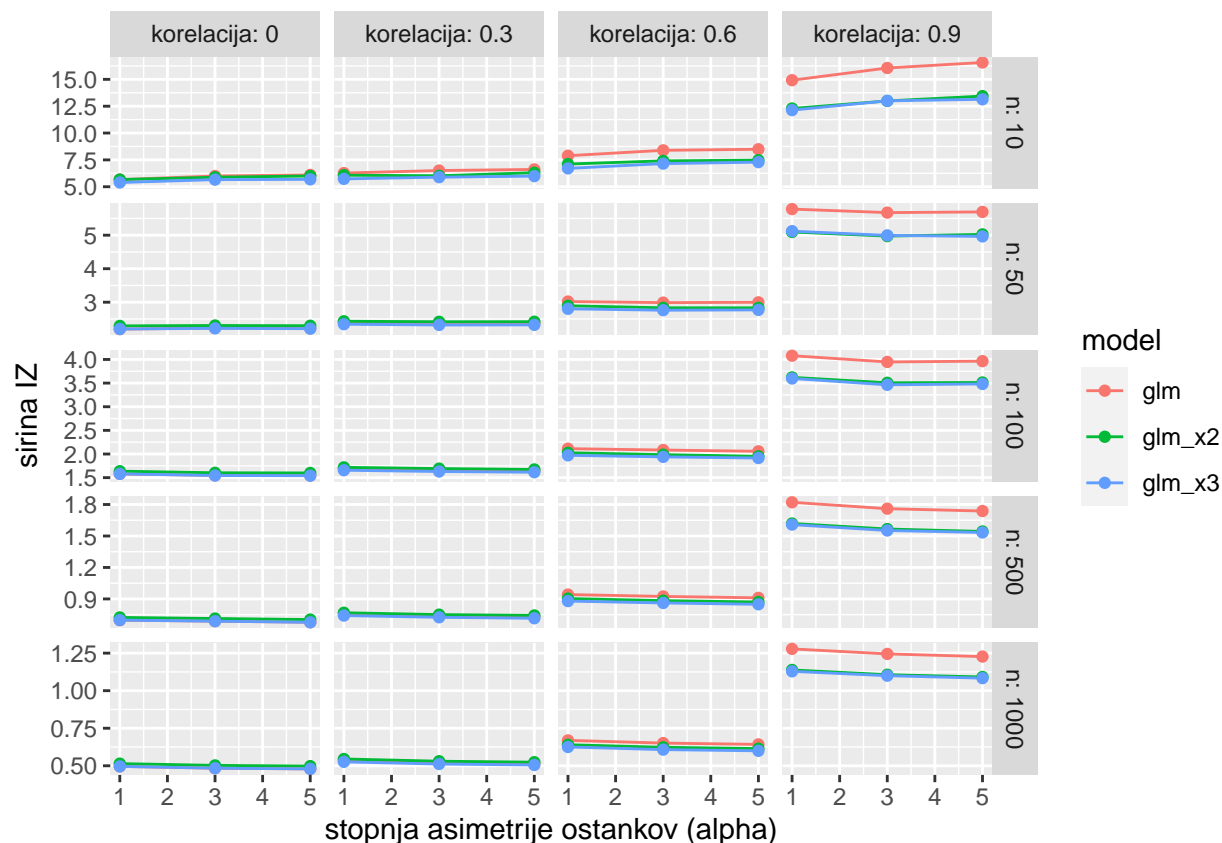
Poglejmo si še vpliv izločanja spremenljivk in ali se ta vpliv razlikuje med posameznimi metodami ter pri različnih porazdelitvah pojasnjevalnih spremenljivk.

Asimetrična porazdelitev pojasnjevalnih spremenljivk

Za začetek si pogledjmo kako se razlikujejo rezultati pri posameznih metodah, potem pa bomo primerjali še posamezne modele z odstranjenimi spremenljivkami pri obeh metodah hkrati.



Slika 6: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk



Slika 7: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk

Pri majhnem vzorcu ni velikih razlik pri pokritosti modela, je pa nekaj razlike v širini intervala zaupanja, predvsem ob prisotnosti korelacije. Odstranitev spremenljivke X_3 nima večjega vpliva, nekoliko višja je pokritost v primeru majhnega vzorca. Močnejše vpliva pa odstranitev spremenljivke X_2 , ki ima dejanski vpliv na odvisno spremenljivko. Najbolj opazen je vpliv v primeru korelacije 0.6, kjer se precej zniža pokritost. V primeru še višje korelacije se pokritost dvigne, kar je pričakovano, saj so spremenljivke že zelo močno povezane in s tem, ko je ne vključimo, povzročimo manj škode. V splošnem so intervali zaupanja v primeru izključene spremenljivke 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. To je najverjetneje posledica tega, da v primeru večjega vzorca nastane tudi večja razlika med pojasnjevalnimi spremenljivkami.

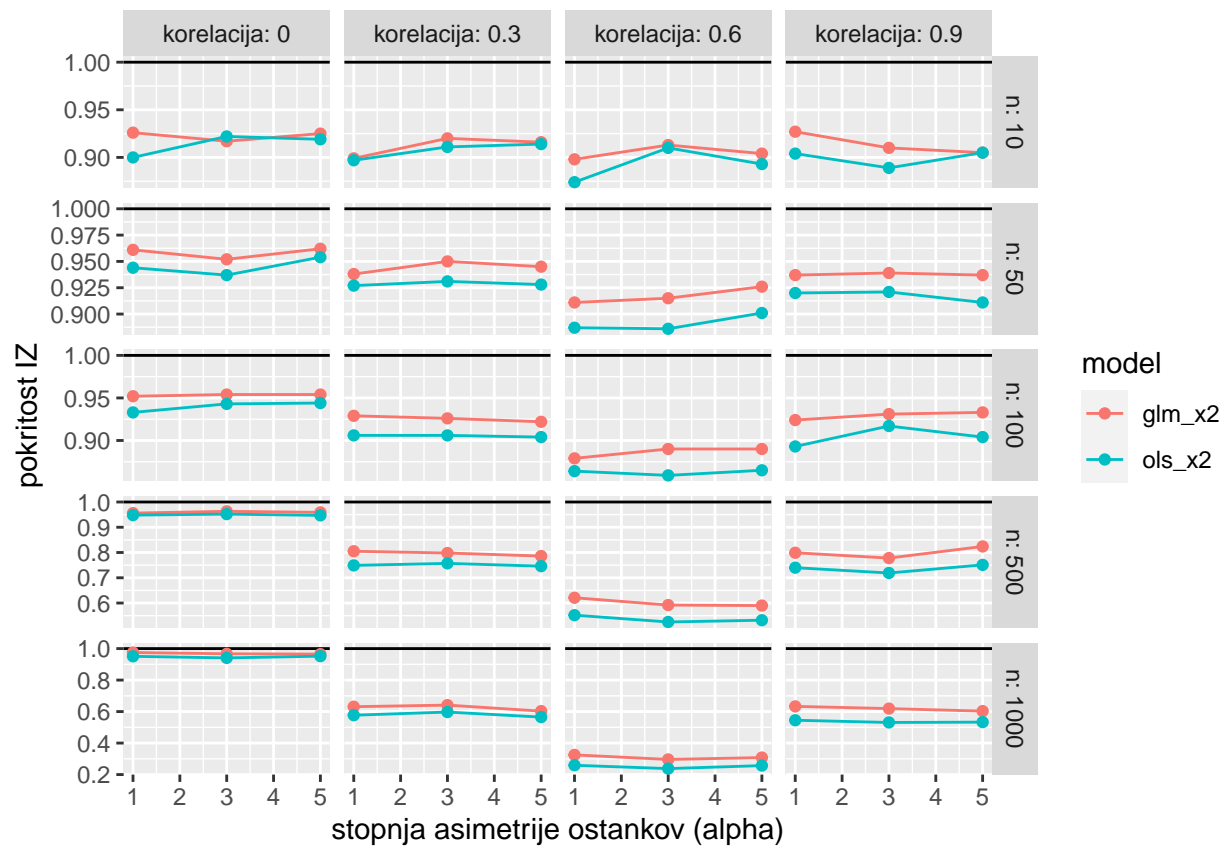
Najboljši model v tem primeru je, ko izključimo X_3 , ki nima vpliva, saj imamo podobno pokritost, a ožji interval zaupanja. V naslednji tabeli si lahko bolje ogledamo dejanske številke.

n	korelacija	glm	glm_x2	glm_x3
10	0.3	0.91	0.91	0.92
10	0.9	0.92	0.91	0.93
100	0.3	0.96	0.93	0.96
100	0.9	0.96	0.93	0.96
1000	0.3	0.96	0.62	0.97
1000	0.9	0.97	0.62	0.97

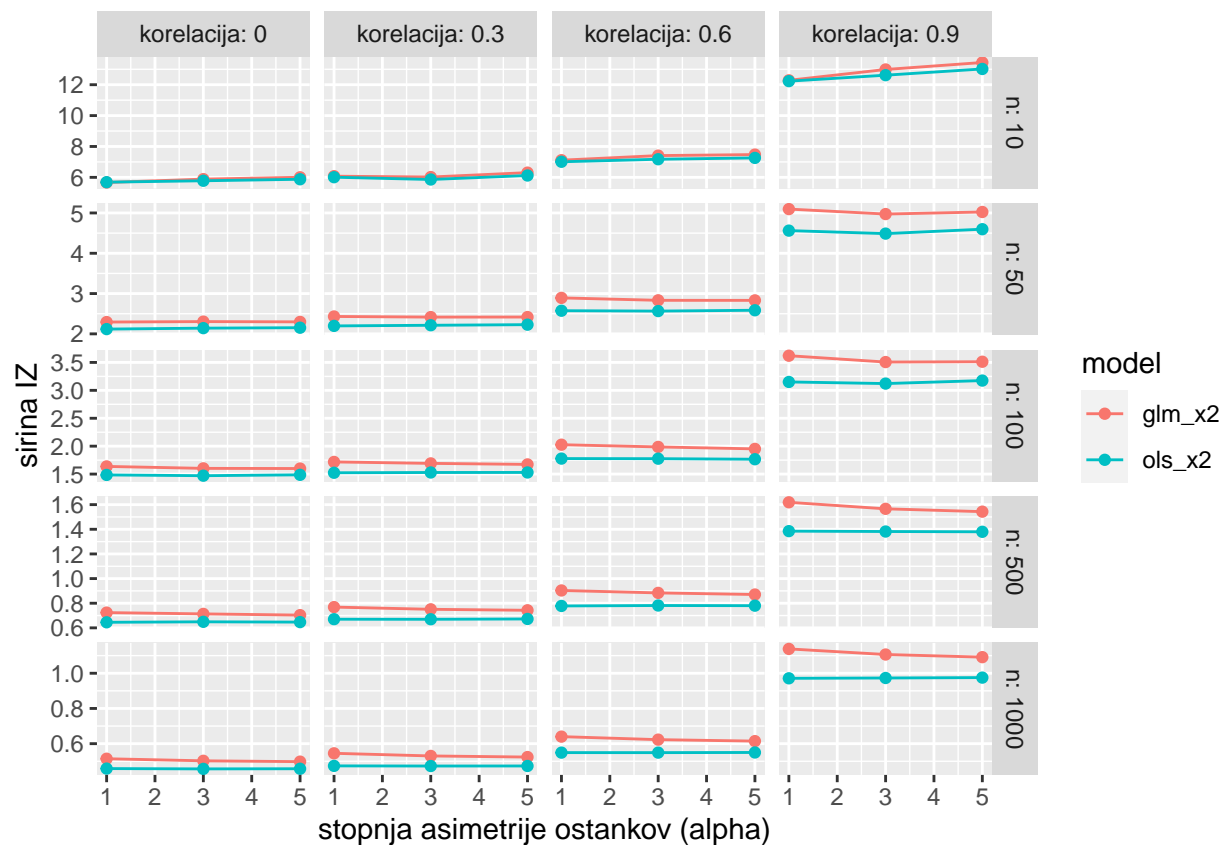
n	korelacija	glm	glm_x2	glm_x3
10	0.3	6.47	6.13	5.87
10	0.9	15.85	12.90	12.76
100	0.3	1.68	1.69	1.63
100	0.9	4.00	3.55	3.51
1000	0.3	0.53	0.53	0.51
1000	0.9	1.25	1.11	1.10

Tabele potrjujejo naša opažanja.

Poglejmo si še razlike med metodama, kjer izločimo eno od spremenljivk. Najprej si pogledjmo rezultate modela brez x_2 .



Slika 8: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez x_2 in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk



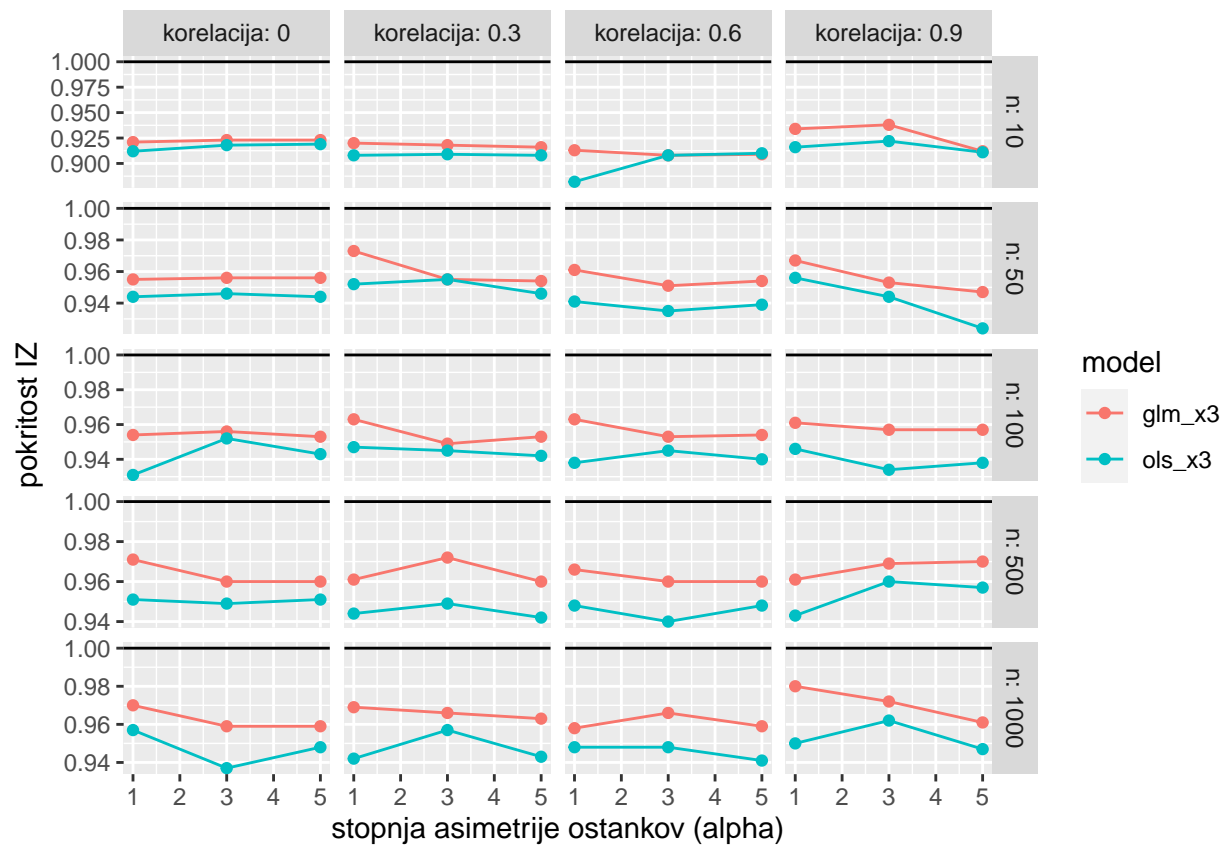
Slika 9: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika
10	0.3	0.91	0.91	0.00
10	0.9	0.91	0.90	0.01
100	0.3	0.93	0.91	0.02
100	0.9	0.93	0.90	0.03
1000	0.3	0.62	0.58	0.04
1000	0.9	0.62	0.54	0.08

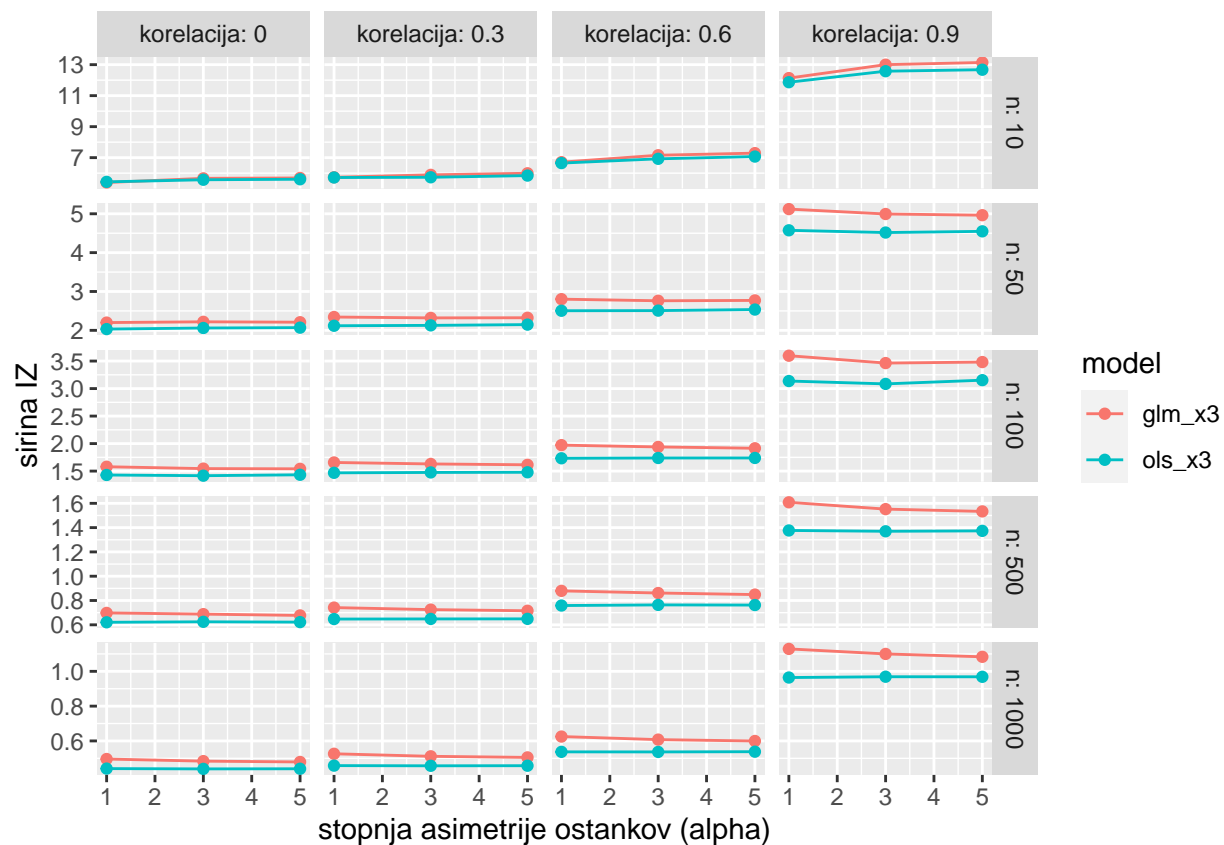
n	korelacija	glm_x2	ols_x2	razlika	razlika_pct
10	0.3	6.13	6.00	0.13	0.02
10	0.9	12.90	12.62	0.28	0.02
100	0.3	1.69	1.53	0.16	0.10
100	0.9	3.55	3.15	0.40	0.13
1000	0.3	0.53	0.47	0.06	0.13
1000	0.9	1.11	0.97	0.14	0.14

Rezultati so podobni kot v primeru polnega modela, večja razlika nastane pri velikih vzorcih. Pri večjih vzorcih glede na polni model nastane večja razlika pri pokritosti. Razlika v širini intervala zaupanja je podobna kot v polnem modelu.

Kaj pa, če odstranimo x_3 ?



Slika 10: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez x_3 in $\text{Gamma}(2,5)$ porazdelitvi pojasnjevalnih spremenljivk



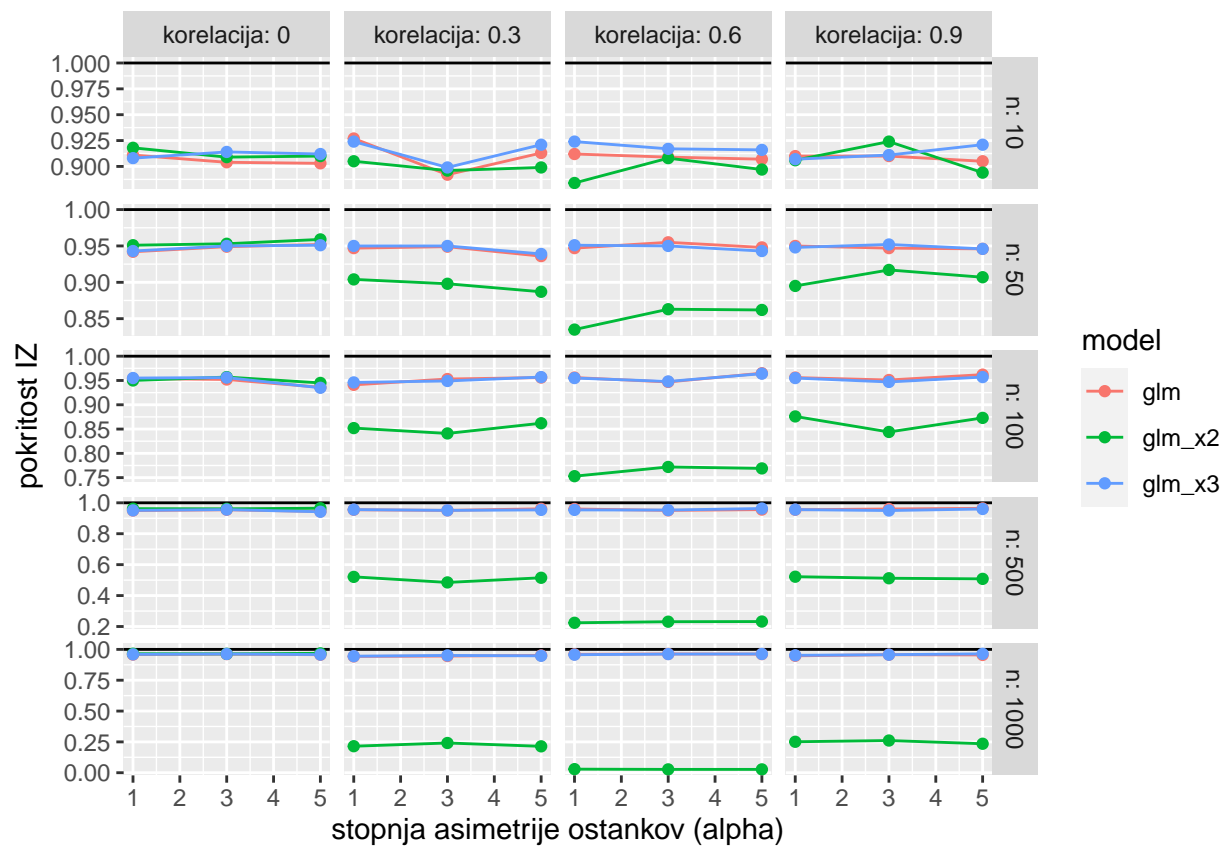
Slika 11: Širina intervalov zaupanja za koeficient β_1 pri modelu brez x3 in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x3	ols_x3	razlika
10	0.3	0.92	0.91	0.01
10	0.9	0.93	0.92	0.01
100	0.3	0.96	0.94	0.02
100	0.9	0.96	0.94	0.02
1000	0.3	0.97	0.95	0.02
1000	0.9	0.97	0.95	0.02

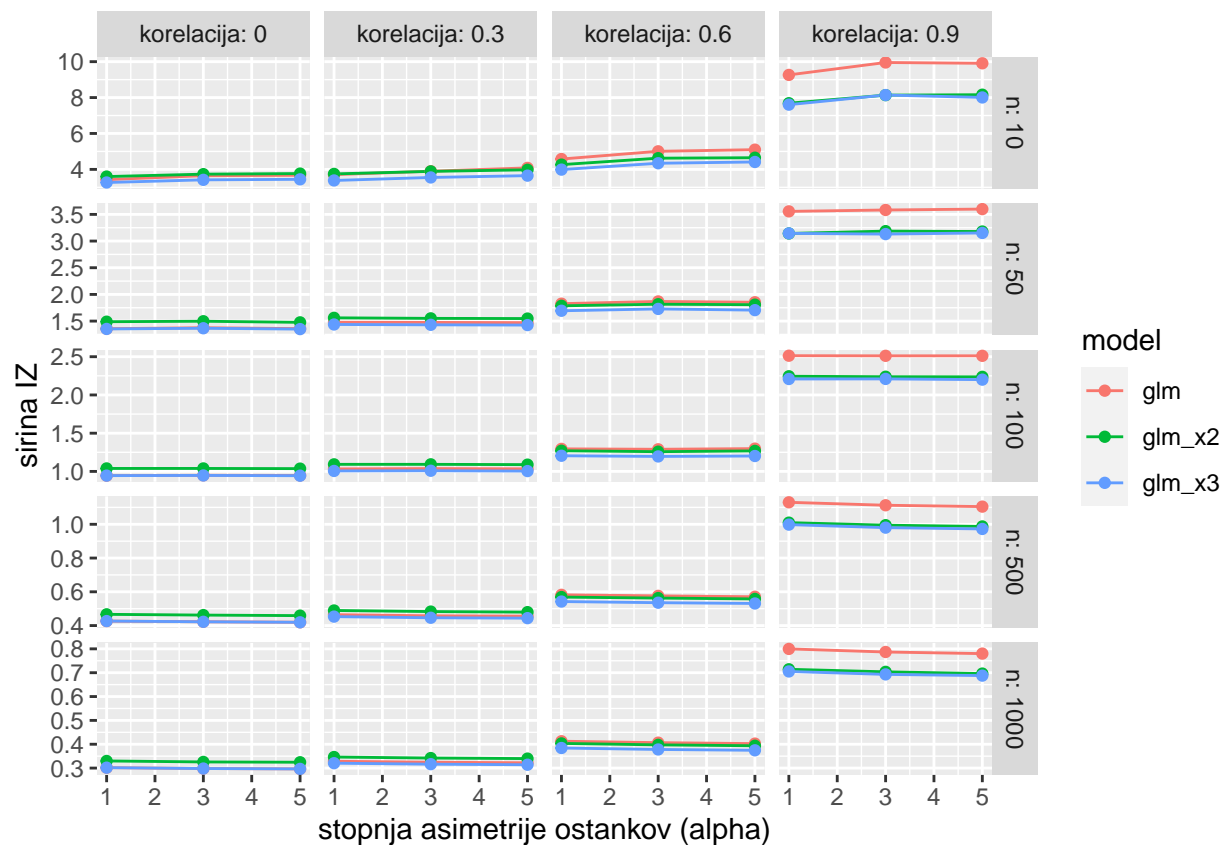
n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
10	0.3	5.87	5.76	0.11	0.02
10	0.9	12.76	12.37	0.39	0.03
100	0.3	1.63	1.47	0.16	0.11
100	0.9	3.51	3.12	0.39	0.12
1000	0.3	0.51	0.46	0.05	0.11
1000	0.9	1.10	0.97	0.13	0.13

Rezultati so zelo podobni kot v primeru polnega modela.

Manj asimetrična porazdelitev pojasnjevalnih spremenljivk



Slika 12: Pokritost intervalov zaupanja za koeficient β_1 pri GLM in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk



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

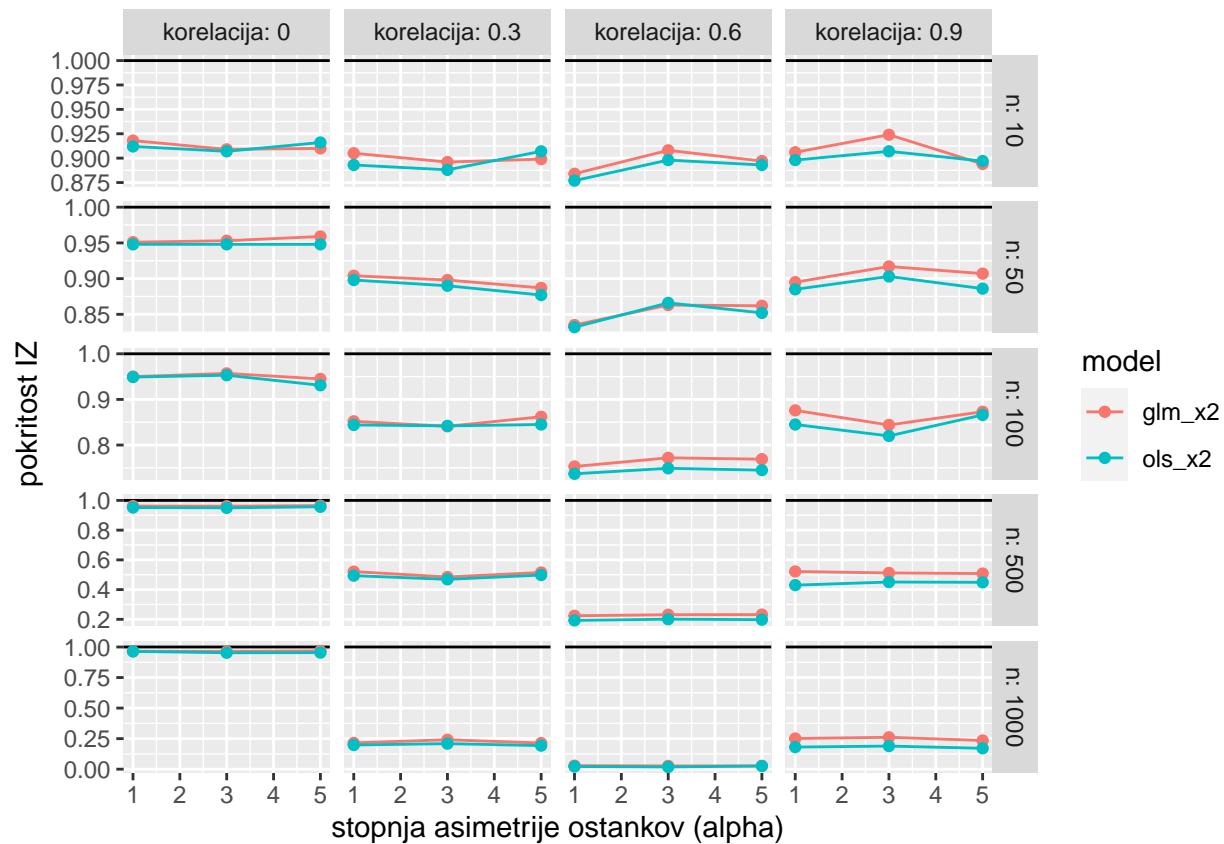
Rezultati so na videz podobni tistim z bolj asimetrično porazdelitvijo pojasnjevalnih spremenljivk.

n	korelacija	glm	glm_x2	glm_x3
10	0.3	0.91	0.90	0.91
10	0.9	0.91	0.91	0.91
100	0.3	0.95	0.85	0.95
100	0.9	0.96	0.86	0.95
1000	0.3	0.95	0.22	0.95
1000	0.9	0.95	0.25	0.96

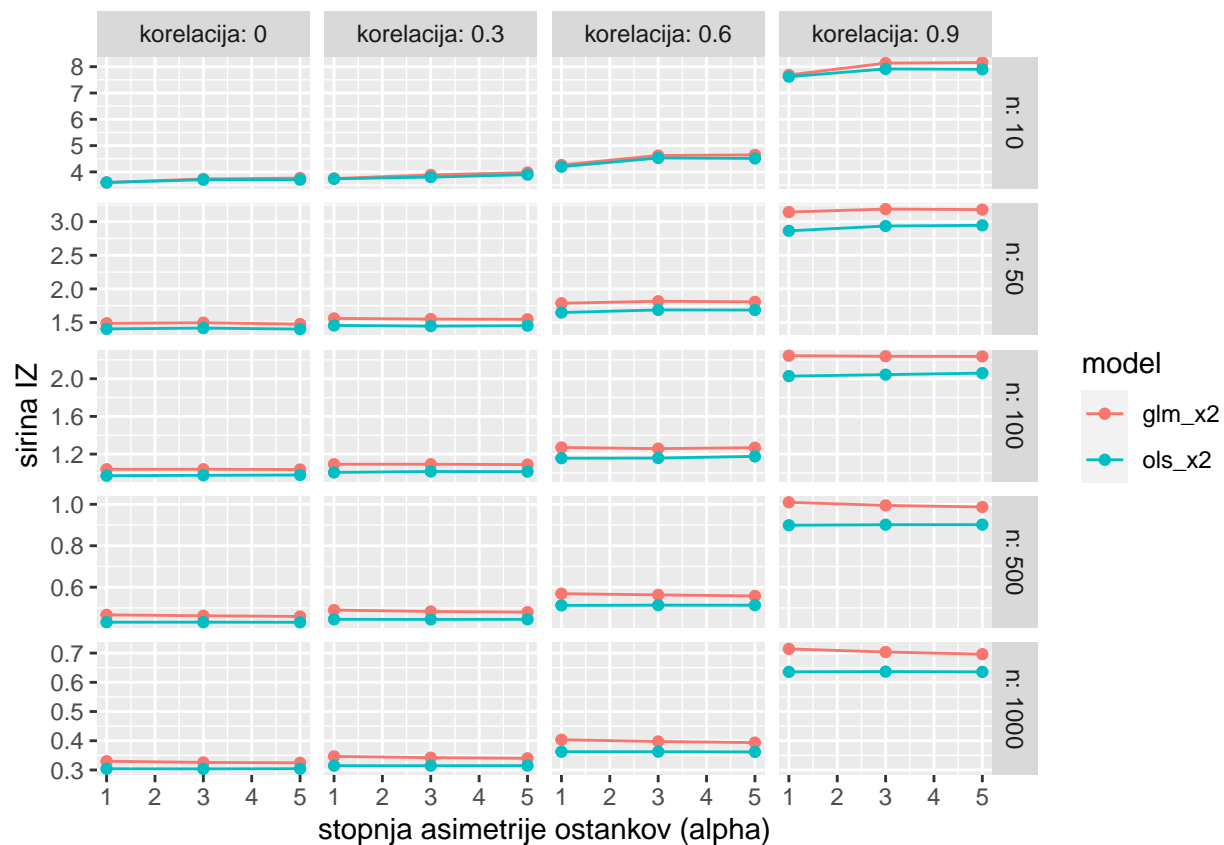
n	korelacija	glm	glm_x2	glm_x3
10	0.3	3.90	3.87	3.53
10	0.9	9.71	7.99	7.92
100	0.3	1.03	1.09	1.01
100	0.9	2.51	2.24	2.21
1000	0.3	0.33	0.34	0.32
1000	0.9	0.79	0.70	0.70

Poglejmo si še razlike med metodama pri odstranjenih spremenljivkah.

Odstranimo x_2 .



Slika 14: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez x_2 in $\text{Gamma}(5,5)$ porazdelitvi pojasnjevalnih spremenljivk



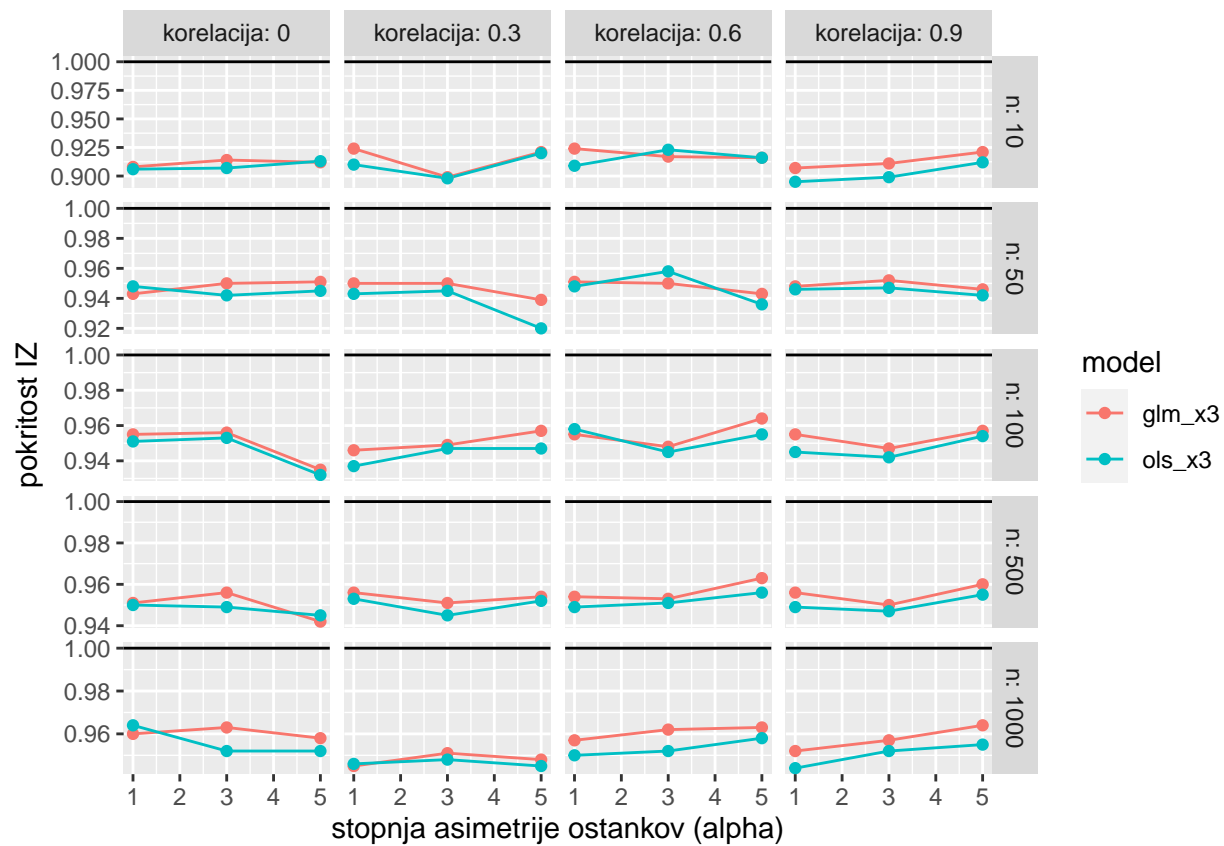
Slika 15: Širina intervalov zaupanja za koeficient β_1 pri polnem modelu in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk

n	korelacija	glm_x2	ols_x2	razlika
10	0.3	0.90	0.90	0.00
10	0.9	0.91	0.90	0.01
100	0.3	0.85	0.84	0.01
100	0.9	0.86	0.84	0.02
1000	0.3	0.22	0.20	0.02
1000	0.9	0.25	0.18	0.07

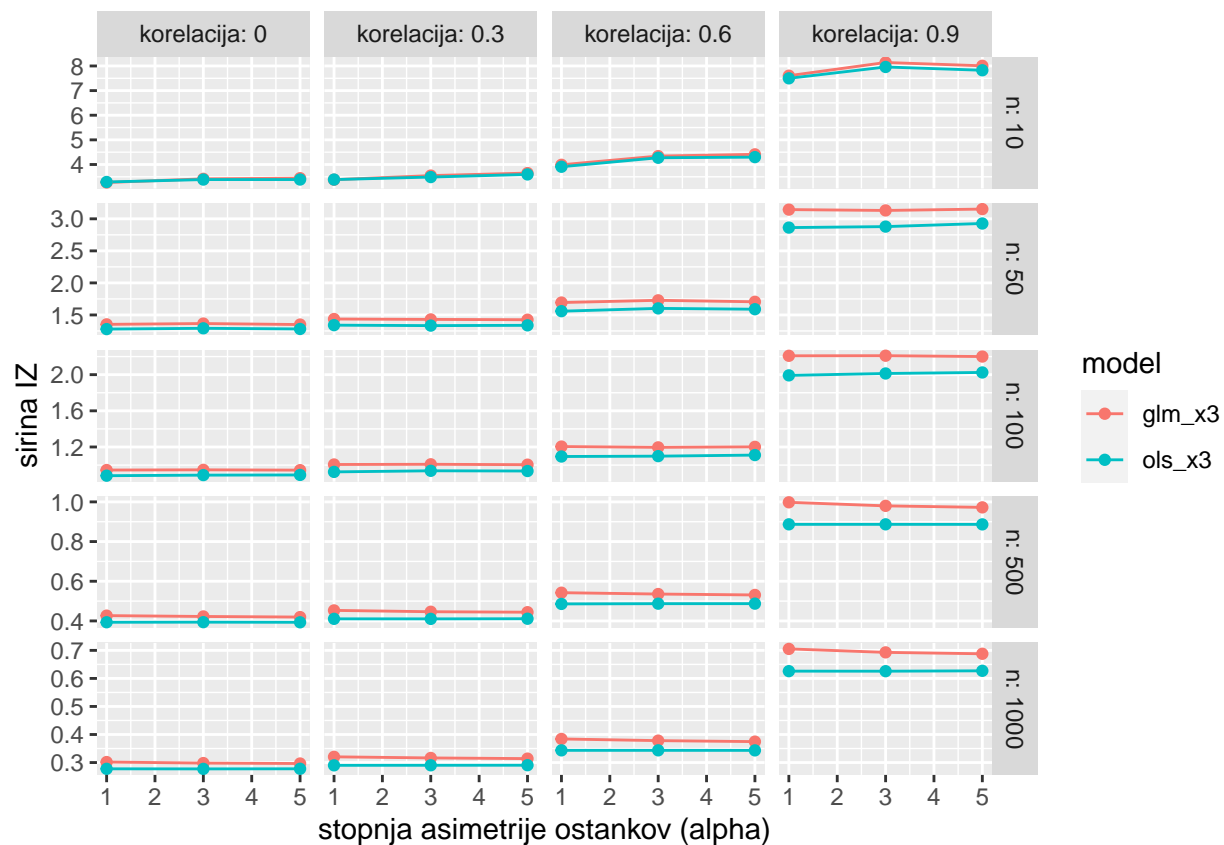
n	korelacija	glm_x2	ols_x2	razlika	razlika_pct
10	0.3	3.87	3.81	0.06	0.02
10	0.9	7.99	7.81	0.18	0.02
100	0.3	1.09	1.01	0.08	0.08
100	0.9	2.24	2.04	0.20	0.10
1000	0.3	0.34	0.31	0.03	0.10
1000	0.9	0.70	0.64	0.06	0.09

Rezultati so podobni kot v primeru polnega modela, vseeno pa je razlika v pokritosti nekoliko višja. Večjo pokritost ima metoda glm, hkrati pa širše intervale zaupanja.

Kaj pa, če odstranimo x_3 ?



Slika 16: Pokritost intervalov zaupanja za koeficient β_1 pri modelu brez x_3 in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk



Slika 17: Širina intervalov zaupanja za koeficient β_1 pri modelu brez x3 in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk

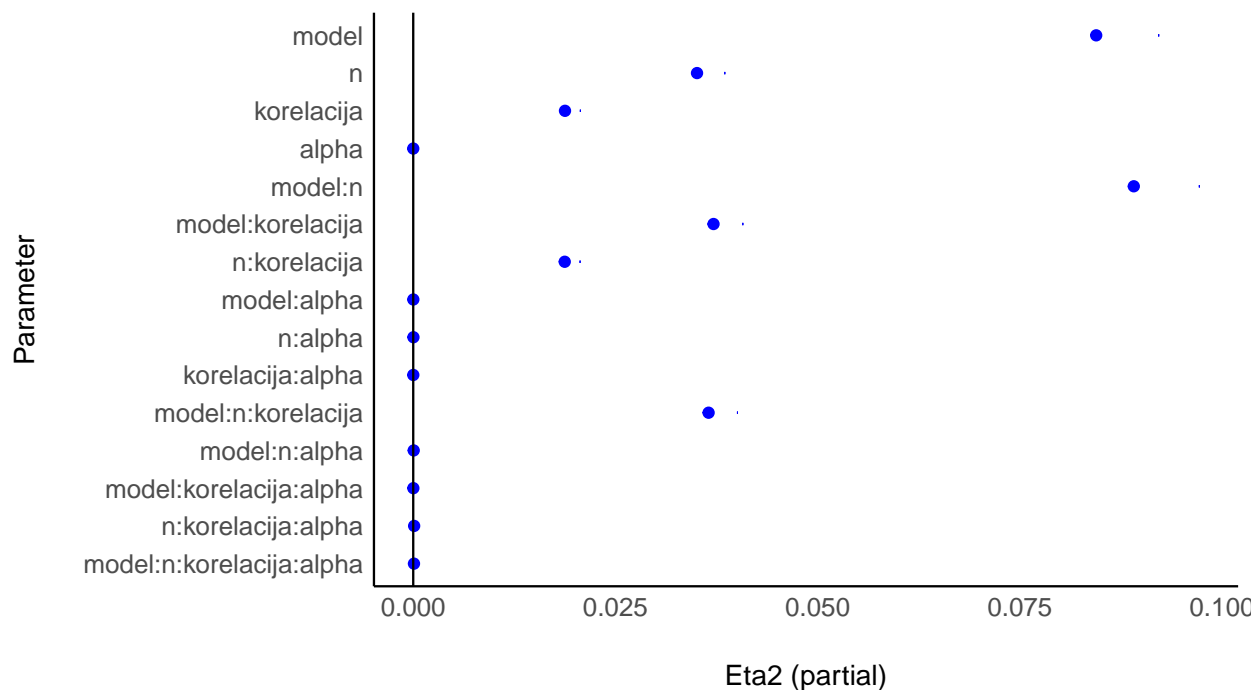
n	korelacija	glm_x3	ols_x3	razlika
10	0.3	0.91	0.91	0.00
10	0.9	0.91	0.90	0.01
100	0.3	0.95	0.94	0.01
100	0.9	0.95	0.95	0.00
1000	0.3	0.95	0.95	0.00
1000	0.9	0.96	0.95	0.01

n	korelacija	glm_x3	ols_x3	razlika	razlika_pct
10	0.3	3.53	3.49	0.04	0.01
10	0.9	7.92	7.76	0.16	0.02
100	0.3	1.01	0.93	0.08	0.09
100	0.9	2.21	2.01	0.20	0.10
1000	0.3	0.32	0.29	0.03	0.10
1000	0.9	0.70	0.63	0.07	0.11

Rezultati so zelo podobni kot v primeru polnega modela.

Analiza variance in velikost učinka

Ker smo že zgoraj ugotovili, da ni večjih razlik pri pokritosti in širini intervalov zaupanja med posameznimi regresijskimi koeficienti, bomo analizo variance naredili 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. Rezultati iz grafov beremo hierarhično, torej koliko posamezna spremenljivka dodatno pojasni variabilnosti, glede na že upoštevane spremenljivke.



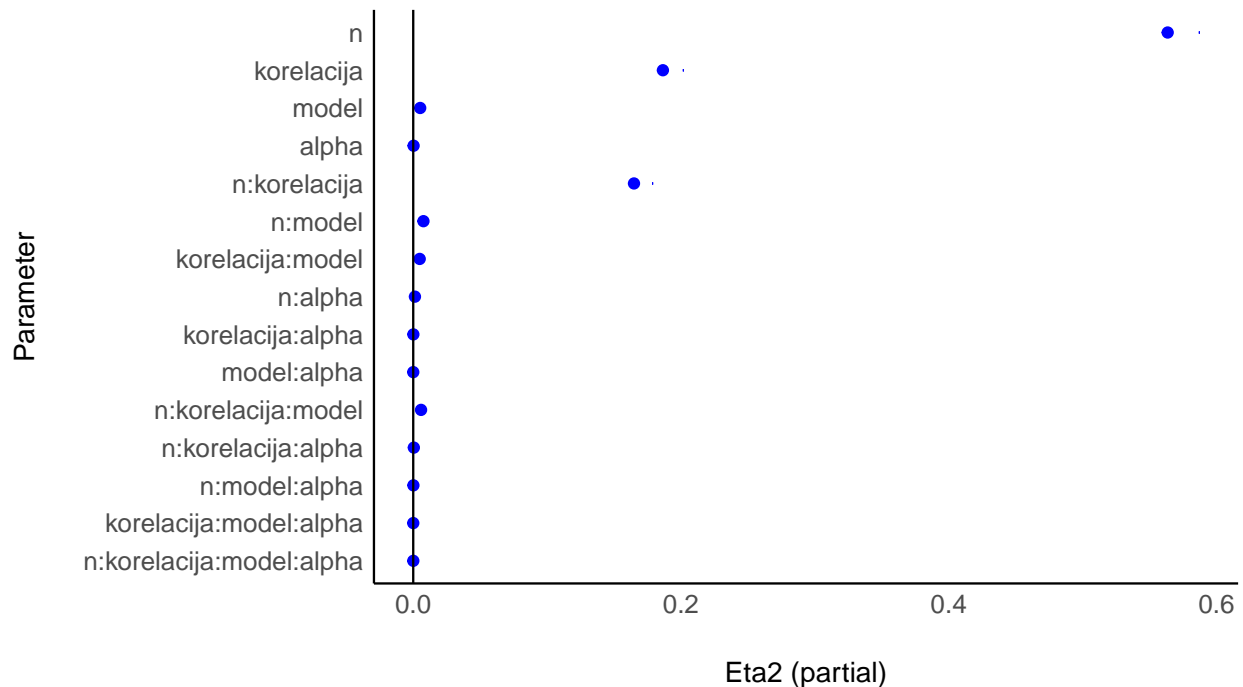
Slika 18: Velikost učinka pri analizi variance za pokritost intervala zaupanja

Iz zgornjega grafa lahko razberemo, da nobena od spremenljivk ne pojasni velikega deleža variabilnosti v pokritosti intervala zaupanja, je pa kar nekaj takih, ki pojasnijo manjši del in ki imajo statistično značilen vpliv (to nam pove povzetek modela anove, ki ni vključen v prikaz).

Vidimo, da na začetku model pojasni nekaj manj kot 10% variabilnosti pokritosti intervala zaupanja. Za tem velikost vzorca pojasni okrog 3% preostale variabilnosti in korelacija nekaj manj kot 2.5%. Asimetričnost ostankov za tem dodatno ne pojasni nič variabilnosti. Za tem nekoliko večji del variabilnosti (glede na ostale vplive) dodatno pojasnijo še interakcija med modelom in velikostjo vzorca, interakcija med modelom in korelacijo ter na koncu še interakcija modela, velikosti vzorca in korelacije.

Pri tem se moramo zavedati, da znotraj spremenljivke model ni samo *glm* ali *lm*, vendar so upoštevani še ne polni modeli. To poudarimo zato, ker je učinek najverjetneje posledica nepolnih modelov, kjer smo tudi v prejšnjih prikazih zaznali vpliv, medtem ko med *lm* in *glm* modelom 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 potrjuje. Poglejmo si še velikost učinka pri analizi variance za širino intervalov zaupanja.



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

Opazimo, da v tem primeru ni tako veliko spremenljivk, ki bi pojasnjevale variabilnost intervalov zaupanja. Okrog 60% variabilnost opazovane spremenljivke pojasni velikost vzorca, od preostale variabilnosti pa nekaj več kot 20% pojasni korelacija med neodvisnimi spremenljivkami. Model in asimetričnost ostankov dodatno ne pojasnita večjega dela variabilnosti širine intervala zaupanja, okrog 20% pa dodatno pojasnjuje še interakcija med velikostjo vzorca in korelacijo. Ponovno nam prikaz in analiza variance potrjujeta naša predhodna opazovanja.

Ugotovitve

Vse za prilagodit novim grafom (brez beta0).

Ugotovili smo, da asimetrija pojasnjevalnih spremenljivk nima pomembnega vpliva na pokritost in širino intervalov zaupanja regresijskih koeficientov. Prav tako smo ugotovili, da med metodama *glm* in *lm* ni posebnih razlik pri ocenah regresijskih koeficientov. Ker naju je ta ugotovitev presenetila, sva poskusili najti podoben problem v kakšni drugi literaturi in naleteli na delo P. E. Johnsona, ki je prišel do podobnih ugotovitev. Sklenil je, da gama porazdelitev ostankov le redkokdaj vpliva na ocene regresijskih koeficientov, kljub temu, da so kršene nekatere predpostavke.

Večja korelacija med pojasnjevalnimi spremenljivkami poveča širino intervalov zaupanja regresijskih koeficientov z izjemo regresijske konstante. Na slednjo korelacija med pojasnjevalnimi spremenljivkami nima posebnega vpliva. Nekoliko negativno vpliva le na pokritost njenega IZ. Na pokritosti intervalov zaupanja ostalih regresijskih koeficientov pa korelacija med pojasnjevalnimi spremenljivkami nima večjega vpliva.

Asimetrija porazdelitve ostankov vpliva le na pokritost IZ regresijske konstante (v resnici je to najbrž posledica pričakovane vrednosti ostankov in ne asimetrije) in v kombinaciji z večjo korelacijo med pojasnjevalnimi spremenljivkami vpliva na širino IZ regresijskih koeficientov.

Z večanjem velikosti vzorca se pričakovano ožajo intervali zaupanja vseh regresijskih koeficientov. Na pokritost IZ velikost vzorca nima vpliva, razen v primeru regresijske konstante. Pokritost IZ regresijske konstante v

primeru gama porazdelitve ostankov nikoli ne doseže velikih vrednosti, saj je kršena predpostavka o ničelni pričakovani vrednosti ostankov.

Izločitev spremenljivke X_3 iz modela po pričakovanjih nima bistvenega vpliva na ocene regresijskih koeficientov, saj smo podatke generirali pod predpostavko $\beta_3 = 0$. Izločitev spremenljivke X_2 pri večjih vzorcih vpliva na slabšo pokritost IZ regresijskega koeficienta β_1 .

Viri

- V. Maver, *Normalni linearni mešani modeli*, diplomsko delo, Fakulteta za matematiko in fiziko, Univerza v Ljubljani, 2018.
- J. Jiang, *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics, Springer Science + Business Media, LLC, New York, 2007.
- *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 02.01.2020], 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.
- 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

Priloge

Vsa uporabljena koda se nahaja v priloženi datoteki `Simulacije.R`.

Sledi izpis ANOVE za prvi in drugi model.

```
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## model              5    5061   1012.2 13274.080 < 2e-16 ***
## n                  4    1995    498.8  6541.821 < 2e-16 ***
## korelacija         3    1049    349.8  4586.991 < 2e-16 ***
## alpha             2         0     0.1     0.744  0.4754
## model:n          20    5367    268.3  3519.030 < 2e-16 ***
## model:korelacija  15    2116    141.1  1849.830 < 2e-16 ***
## n:korelacija     12    1047     87.3  1144.648 < 2e-16 ***
## model:alpha      10         1     0.1     0.690  0.7345
## n:alpha          8         1     0.2     2.095  0.0327 *
## korelacija:alpha  6         0     0.0     0.615  0.7186
## model:n:korelacija 60    2080    34.7   454.580 < 2e-16 ***
## model:n:alpha    40         3     0.1     1.013  0.4472
## model:korelacija:alpha 30         1     0.0     0.329  0.9998
## n:korelacija:alpha 24         6     0.3     3.393 3.6e-08 ***
## model:n:korelacija:alpha 120         5     0.0     0.539  1.0000
## Residuals      719640   54876     0.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## n	4	3572142	893036	2.323e+05	< 2e-16 ***
## korelacija	3	634219	211406	5.499e+04	< 2e-16 ***
## model	5	14450	2890	7.518e+02	< 2e-16 ***
## alpha	2	728	364	9.473e+01	< 2e-16 ***
## n:korelacija	12	546242	45520	1.184e+04	< 2e-16 ***
## n:model	20	21230	1061	2.761e+02	< 2e-16 ***
## korelacija:model	15	13391	893	2.322e+02	< 2e-16 ***
## n:alpha	8	3521	440	1.145e+02	< 2e-16 ***
## korelacija:alpha	6	235	39	1.020e+01	2.54e-11 ***
## model:alpha	10	41	4	1.065e+00	0.385
## n:korelacija:model	60	16229	270	7.036e+01	< 2e-16 ***
## n:korelacija:alpha	24	1239	52	1.343e+01	< 2e-16 ***
## n:model:alpha	40	322	8	2.097e+00	6.00e-05 ***
## korelacija:model:alpha	30	25	1	2.170e-01	1.000
## n:korelacija:model:alpha	120	123	1	2.680e-01	1.000
## Residuals	719640	2766502	4		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1