Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2021-01-10

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.

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 odvisna spremenljivka y tudi normalno porazdeljena s.s.: $y_i \sim N(\beta_1 x_{1i} + ... + \beta_k x_{ki}, \sigma_u^2)$

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:

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

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: $X_j \sim Gamma(\delta, 5)$, $j = 1, 2, 3$, $\delta = 2, 5$;
- 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

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

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.

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. 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 standrdnih 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 ter slabšo pokritost ne glede na izbiro metode.

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.

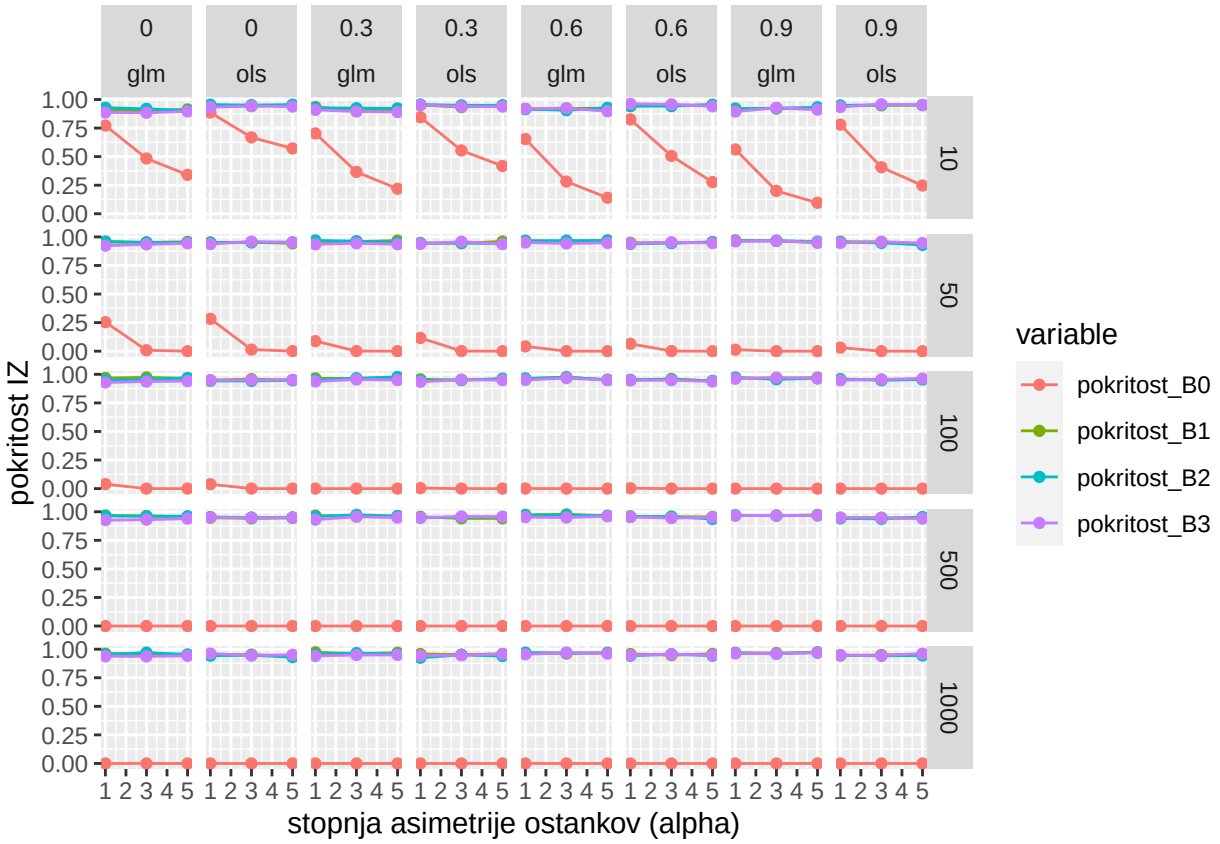
Pričakujemo, da lahko kršenje predpostavke o normalni porazdeljenosti ostankov rešimo z uporabo posplošenih linearnih modelov z ustrezno definirano porazdelitvijo ostankov oz. odzivne spremenljivke. Pričakujemo, da bolj kot bo porazdelitev ostankov asimetrična (manjša vrednost parametraα), slabši bodo 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 - $gama(2,5)$, ki je precej asimetrična in $gama(5,5)$, 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.

Predstavitev rezultatov

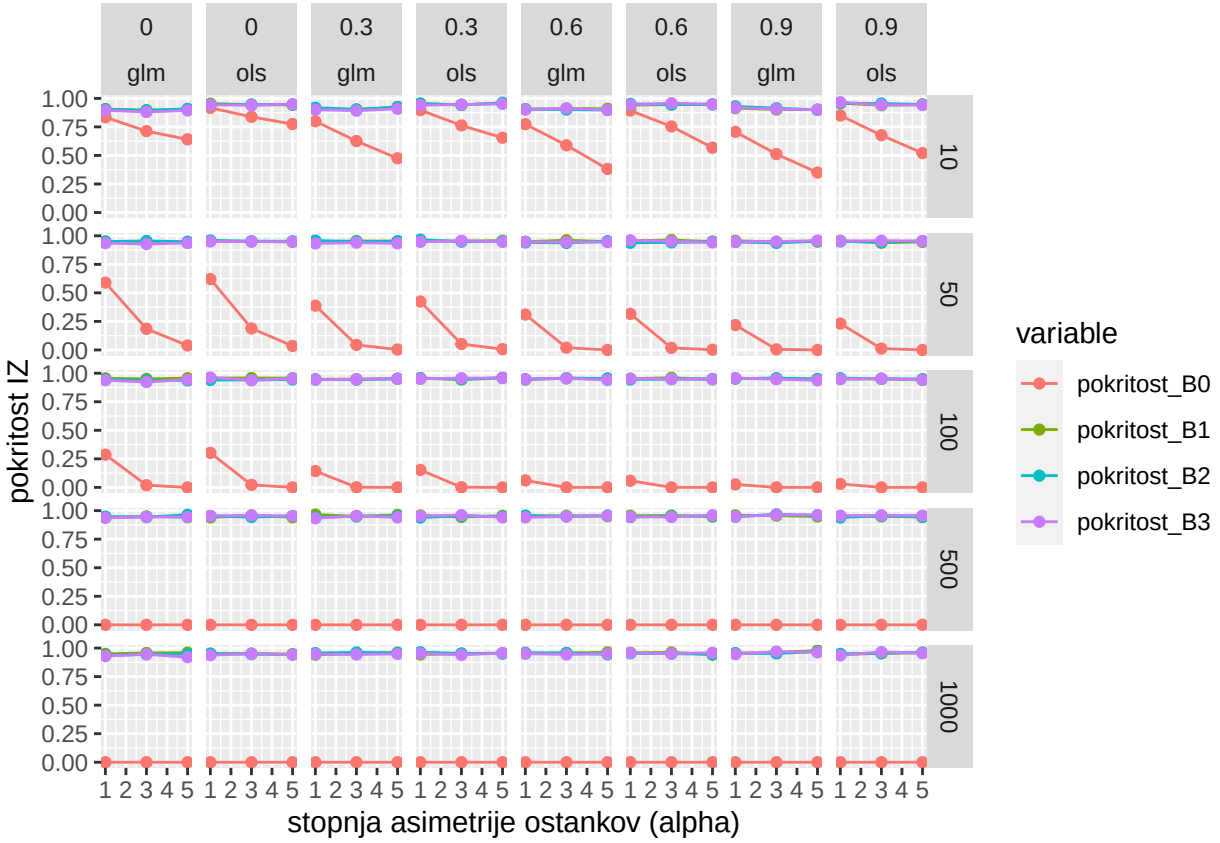
Pokritost intervalov zaupanja



Slika 1: Pokritost intervalov zaupanja pri polnem modelu in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk

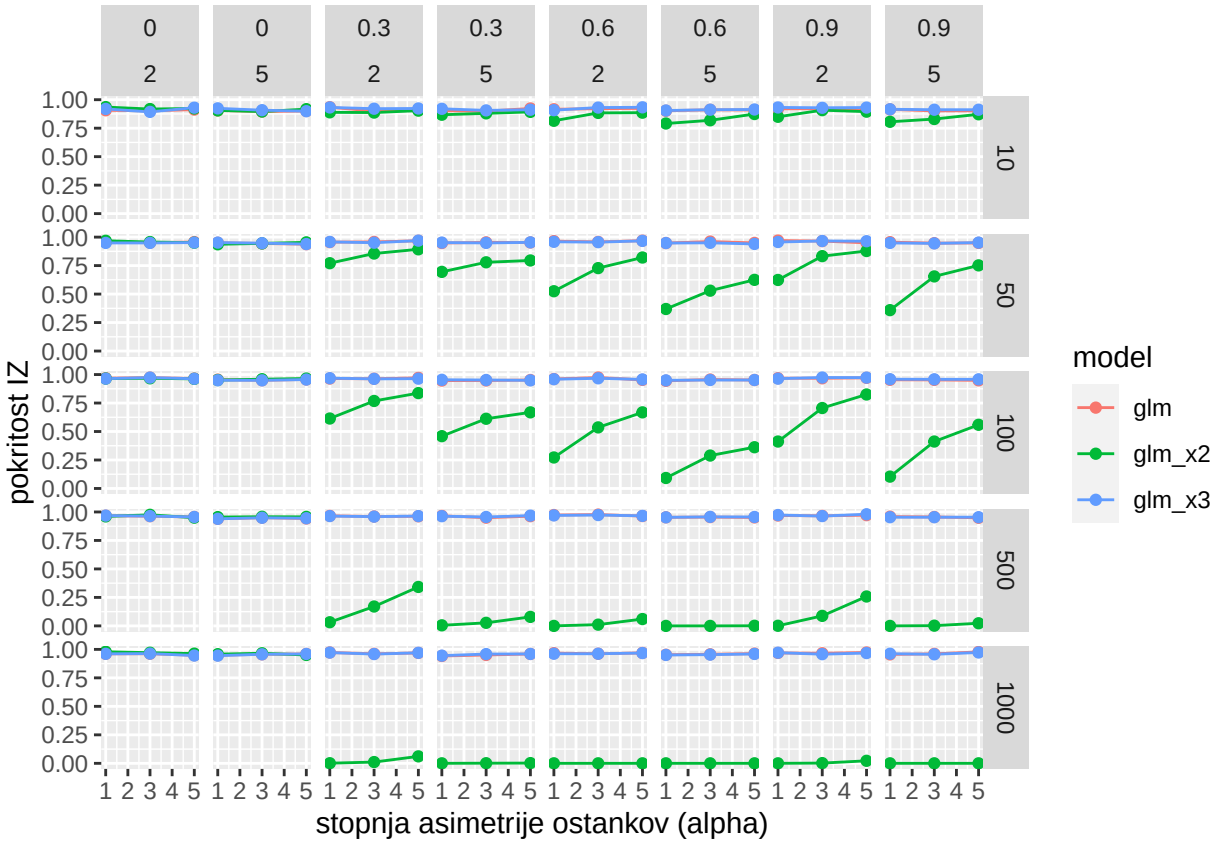
Na Sliki 1 vidimo, da so pokritosti intervalov zaupanja za regresijske koeficiente β_1, β_2 in β_3 vedno blizu 100%, ne glede na uporabljeno metodo, velikost vzorca, korelacijo med pojasnjevalnimi spremenljivkami in stopnjo asimetrije ostankov. Močno izstopa pokritost konstante. Pri večjem vzorcu je njena pokritost praktično 0, pri manjših vzorcih pa se opazi vpliv asimtrije 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 spremenljvkami prav tako vpliva na pokritost koeficienta $beta_0$ - večja kot je korelacija, slabša je pokritost.



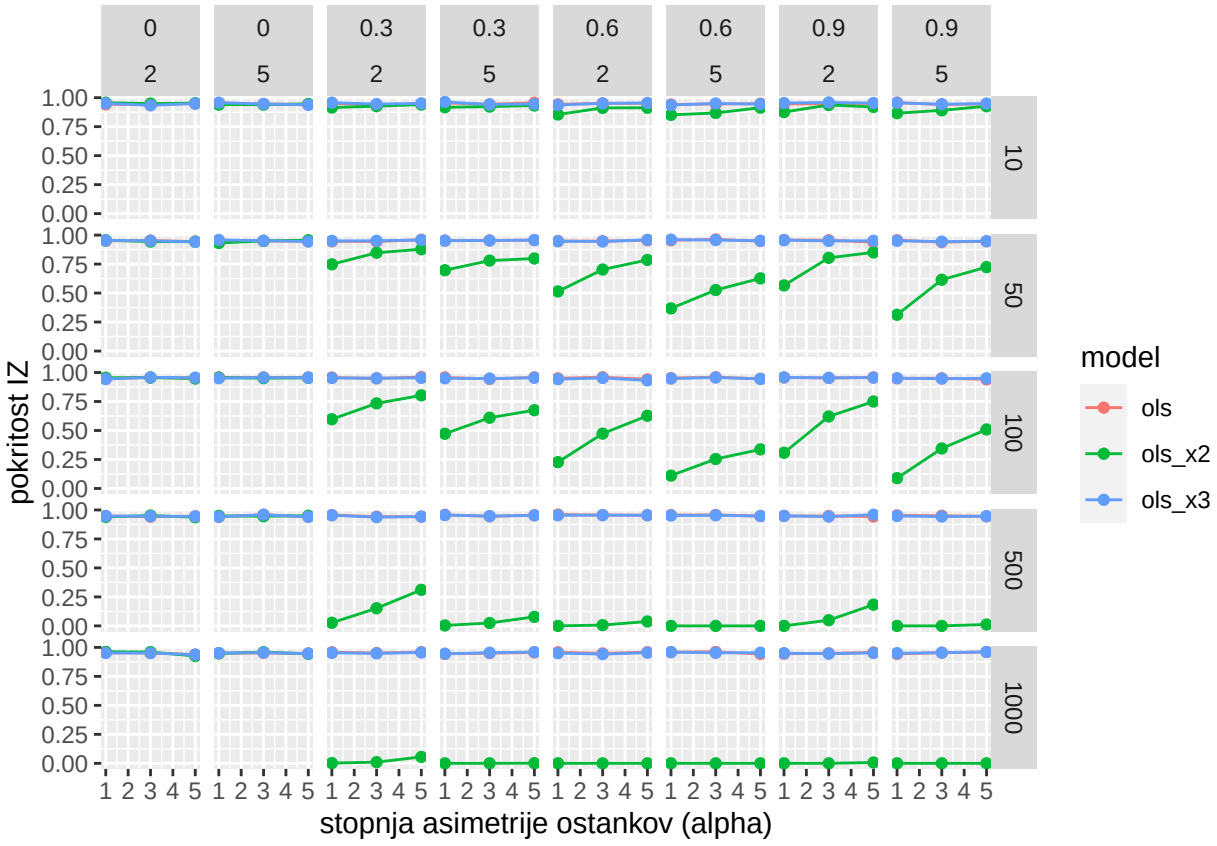
Slika 2: Pokritost intervalov zaupanja pri polnem modelu in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk

Če primerjamo sliki 1 in 2, vidimo, da pri simetrični in asimetrično porazdelitvi pojasnjevalnih spremenljivk dobimo praktično identične rezultate pokritosti intervalov zaupanja regresijskih koeficientov. Interpretacija je torej enaka kot pri Sliki 1.



Slika 3: Pokritost intervalov zaupanja za B1 GLM

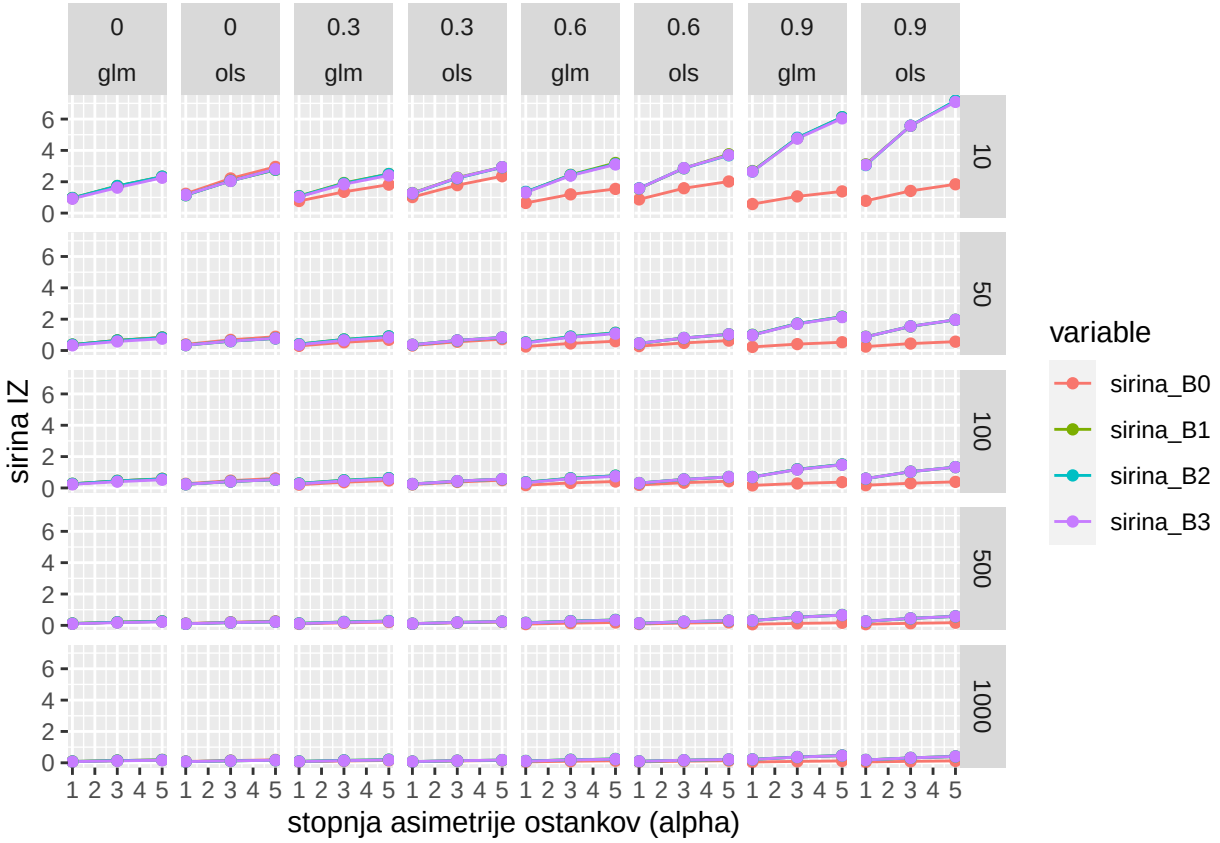
Na zgornjem grafu opazujemo pokritost intervalov zaupanja regresijskega koeficienta β_1 pri polnem modelu (*glm*), pri modelu, v katerega ne vključimo spremenljivke X_2 (*glm_x2*) ter pri modelu, v katerega ne vključimo spremenljivke X_3 (*glm_x3*). Če med pojasnjevalnimi spremenljivkami ni korelacije, je ne glede na ostale dejavnike pokritost IZ vedno blizu 100%. Opazimo, da je pri večji korelaciji med pojasnjevalnimi spremenljivkami in pri večjem vzorcu v modelu brez spremenljivke X_2 manjša pokritost IZ regresijskega koeficienta β_1. Pri polnem modelu in modelu brez spremenljivke X_3 noben dejavnik ne vpliva na pokritost IZ, saj je povsod zelo blizu 100%. Zanimivo je, da je v modelu brez X_2 pokritost pri simetrični porazdelitvi pojasnjevalnih spremenljivk ($gamma(5,5)$) slabša kot pri bolj asimetrični ($gamma(2,5)$). Stopnja asimetrije porazdelitve napak ne vpliva na pokritost IZ regresijskega koeficienta β_1.



Slika 4: Pokritost intervalov zaupanja za B1 OLS

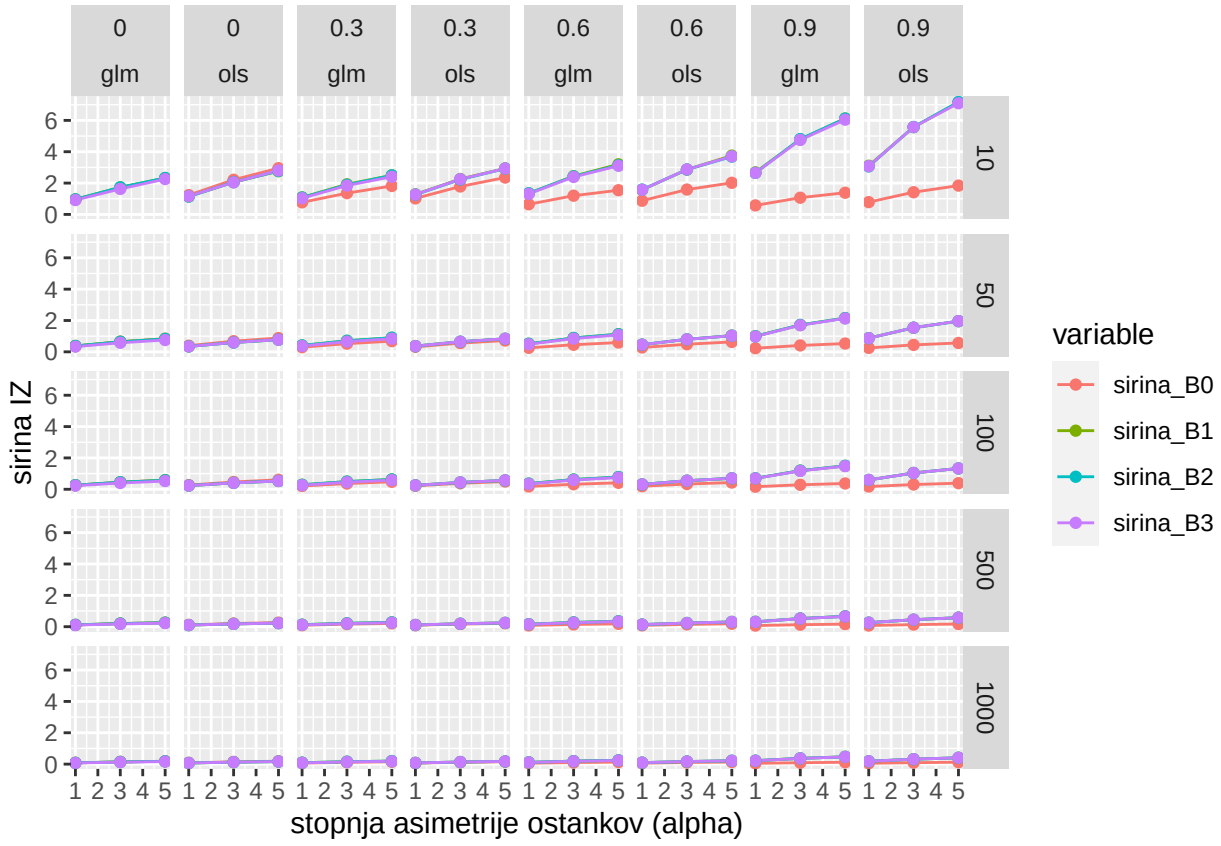
Slika 4 predstavlja isto kot Slika 3, le da je namesto funkcije *glm* uporabljena funkcija *lm*. Opazimo, da med *lm* in *glm* ni bistvenihi razlik, zato interpretacije rezultatov ne bomo ponavljali. Z obema funkcijama dobimo praktično identično pokritost intervalov zaupanja regresijskih koeficientov (konkretno za β_1).

Širina intervalov zaupanja



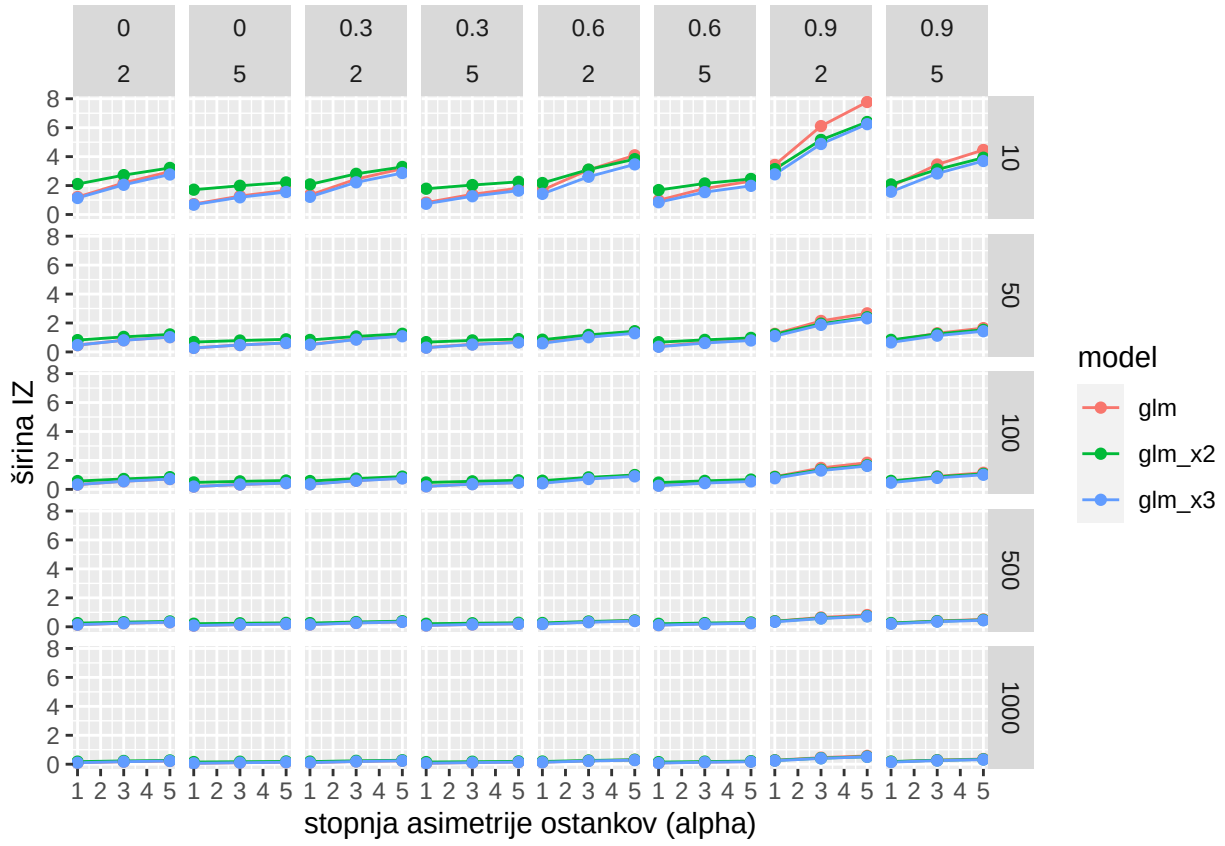
Slika 5: Širina intervalov zaupanja pri polnih modelih in Gamma(2,5) porazdelitvi pojasnjevalnih spremenljivk

Na zgornji sliki vidimo, da če pojasnjevalne spremenljivke med seboj niso korelirane, se širine IZ vseh regresijskih koeficientov (tudi konstante) praktično povsem ujemajo. Pričakovano je pri manjšem vzorcu večja širina kot pri večjem vzorcu. Pri večji korelaciji med pojasnjevalnimi spremenljivkami prihaja do večjih razlik med širinami IZ regresijske konstante in drugih regresijskih koeficientov. Največja širina IZ regresijskih koeficientov je pri najmanjšem vzorcu in največji korelaciji med pojasnjevalnimi spremenljivkami, kar je bilo pričakovano. Pri večjih vzorcih so razlike manjše, vseeno pa tudi pri največjem vzorcu ($n = 1000$) opaziti nekoliko ožji IZ regresijske konstante kot IZ ostalih regresijskih koeficientov. Korelacija med pojasnjevalnimi spremenljivkami na širino IZ regresijske konstante torej manj vpliva. Med metodama *gls* in *lm* v nobenem primeru (ne glede na ostale dejavnike) ni opaznih večjih razlik.



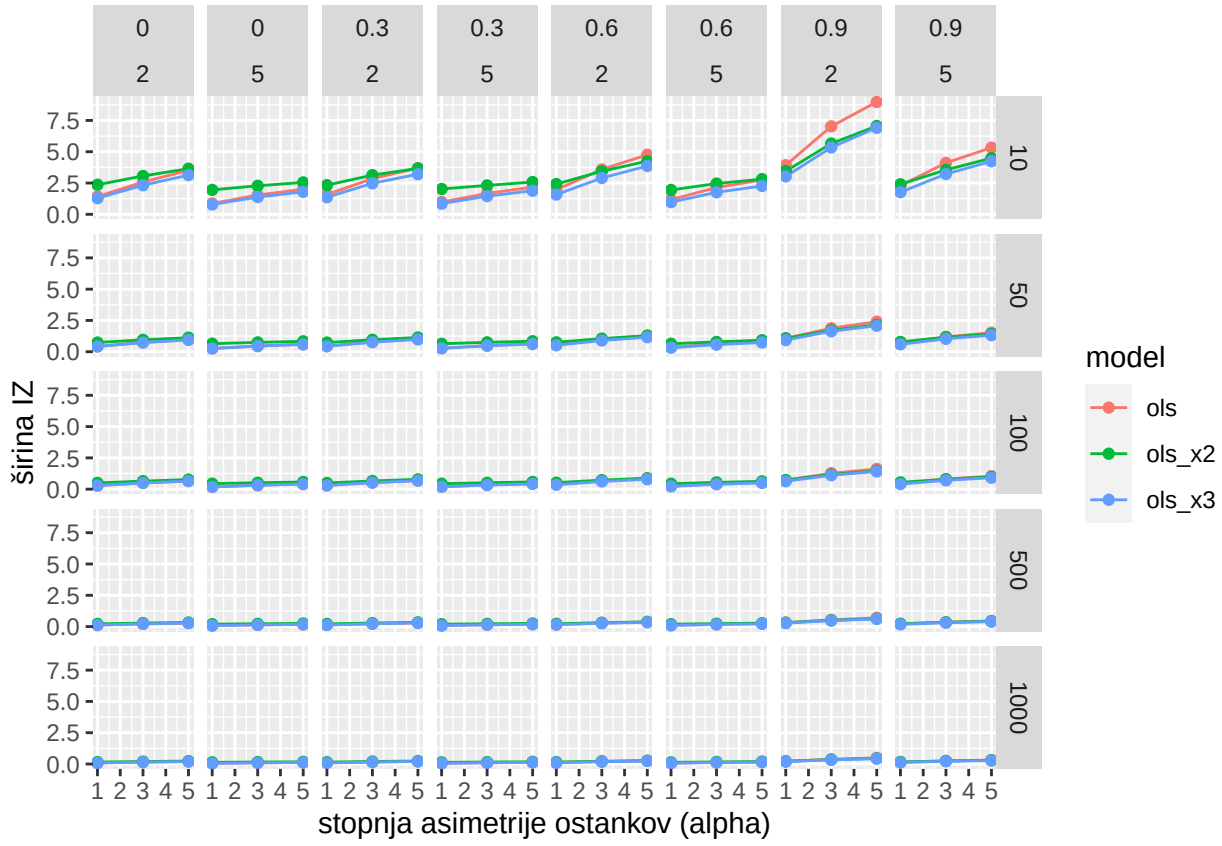
Slika 6: Širina intervalov zaupanja pri polnih modelih in Gamma(5,5) porazdelitvi pojasnjevalnih spremenljivk

Podobno kot pri pokritosti IZ lahko tudi pri širini IZ vidimo, da med slikama 3 in 4 ni posebnih razlik. Torej ni razlik v širini IZ regresijskih koeficientov glede na (a)simetrijo pojasnjevalnih spremenljivk.



Slika 7: Sirina intervalov zaupanja za B1 GLM

Na Sliki 5 so prikazane širine IZ regresijskega koeficienta β_1 z uporabo funkcije *glm* pri polnem modelu, pri modelu brez X_2 (*glm_x2*) in pri modelu brez X_3 (*glm_x3*). Opazimo, da je pri bolj asimetrični porazdelitvi pojasnjevalnih spremenljivk ($gama(2,5)$) IZ širši kot pri bolj simetrični porazdelitvi pojasnjevalnih spremenljivk ($gama(5,5)$). Razlika je bolj opazna pri manjših vzorcih, pri večjih pa se razlike ne opazi več. Stopnja asimetrije porazdelitve ostankov ne vpliva na širino IZ regresijskega koeficienta β_1. Daleč najširše intervale zaupanja dobimo pri manjših vzorcih in veliki korelaciji med pojasnjevalnimi spremenljivkami. Če iz modela izločimo kakšno spremenljivko, to očitno nima nima vpliva na širino IZ regresijskega koeficienta β_1, razen v primeru zelo majhnega vzorca in velike korelacije med pojasnjevalnimi spremenljivkami. V teh primerih dobimo s polnim modelom nekoliko presenetljivo širše intervale zaupanja. Pri izločitvi spremenljivke X_2 smo morda pričakovali slabše rezultate, vendat pa večja korelacija med pojasnjevalnimi spremenljivkami lahko pojasni rezultate.

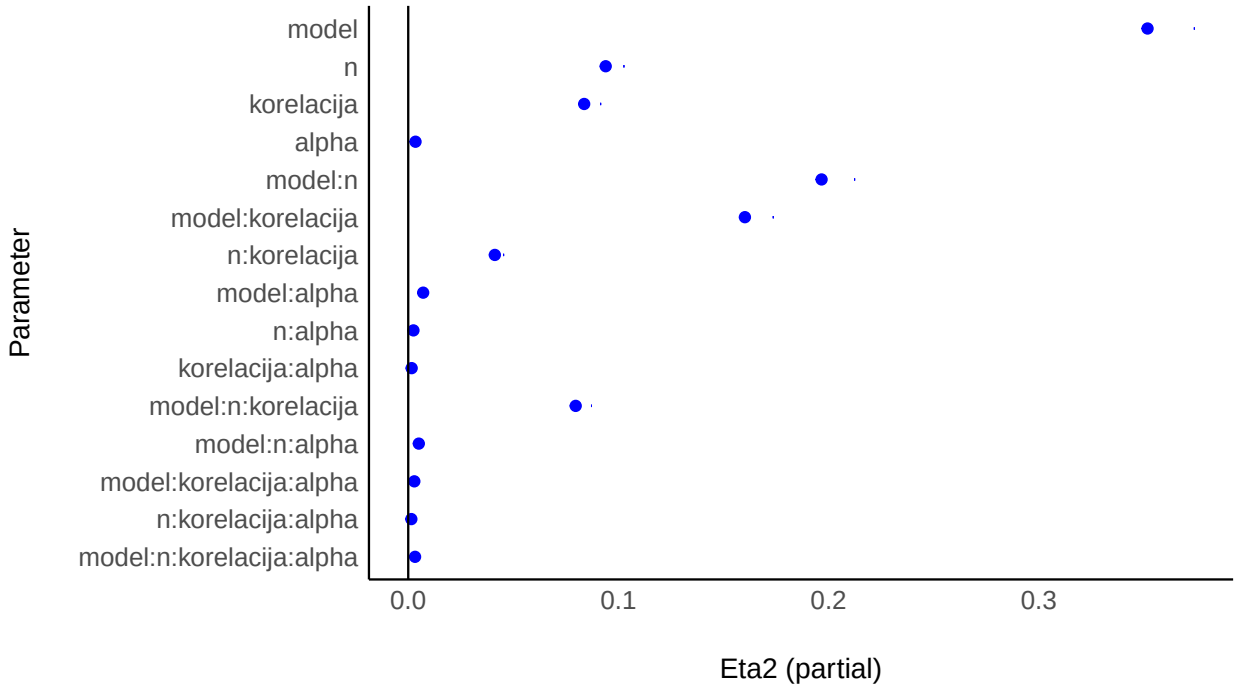


Slika 8: Sirina intervalov zaupanja za B1 LM

Podobno kot pri prejšnjih primerih ugotovimo, da izbira metode (*glm* ali *lm*) ne vpliva na širino IZ regresijskega koeficienta β_1.

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.



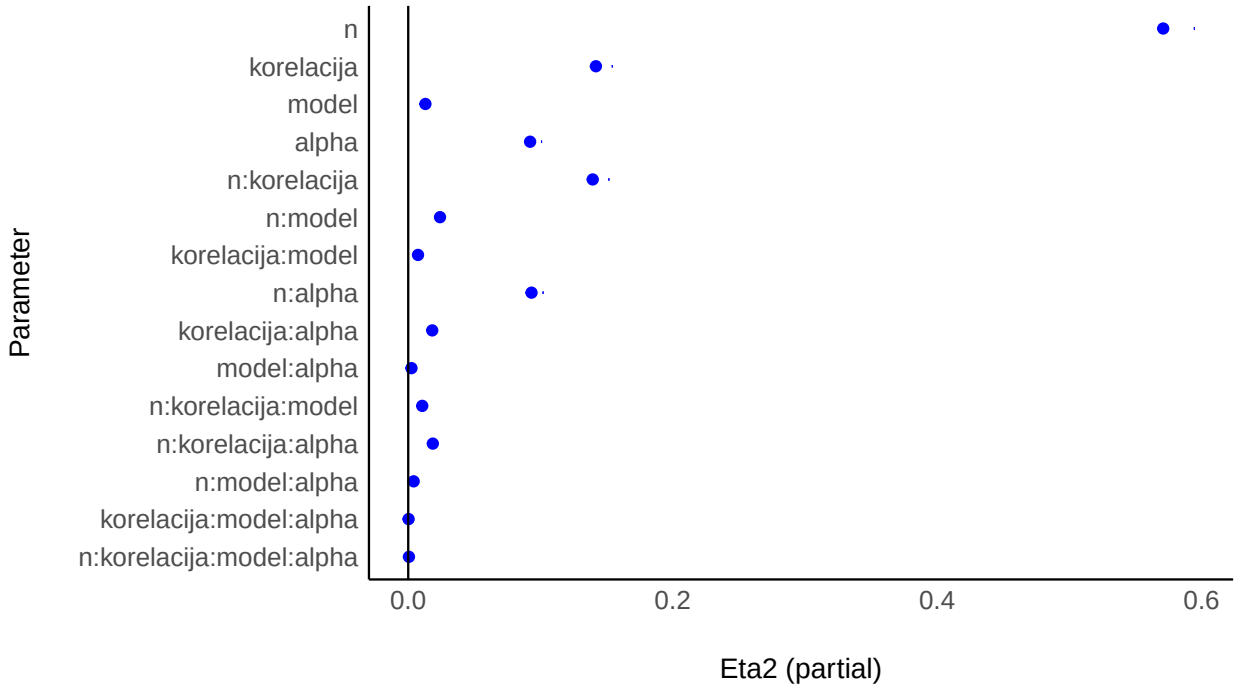
Slika 9: Velikost učinka pri analizi variance za pokritost intervala zaupanja

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 vljučen 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 potrdijo. Poglejmo si še velikost učinka pri analizi variance za širino intervalov zaupanja.



Slika 10: 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 opazovanje spremenljivke pojasni velikost vzorca, od preostale variabilnosti nekaj več kot 20% pojasni korelacija med neodvisnimi spremenljivkami. Model in asimetričnost ostankov dodatno ne pojasnjujeta 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 prejšnja opazovanja.

Ugotovitve

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 spremenljivkai 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 izjemo regresijske konstante).

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 regresijji*, 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  24896    4979 78082.73 <2e-16 ***
## n                             4   4758    1189 18652.22 <2e-16 ***
## korelacija                    3   4192    1397 21911.54 <2e-16 ***
## alpha                         2    159      80  1248.16 <2e-16 ***
## model:n                      20  11232     562  8807.12 <2e-16 ***
## model:korelacija             15   8752     583  9149.54 <2e-16 ***
## n:korelacija                 12   1972     164  2576.61 <2e-16 ***
## model:alpha                  10    328      33   514.85 <2e-16 ***
## n:alpha                       8    114      14   222.62 <2e-16 ***
## korelacija:alpha              6     75      12   194.80 <2e-16 ***
## model:n:korelacija           60   3972      66  1038.15 <2e-16 ***
## model:n:alpha                40    232       6    90.95 <2e-16 ***
## model:korelacija:alpha       30    133       4    69.57 <2e-16 ***
## n:korelacija:alpha           24     68       3    44.33 <2e-16 ***
## model:n:korelacija:alpha    120    150       1    19.54 <2e-16 ***
## Residuals                719640  45890       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                             Df Sum Sq Mean Sq   F value Pr(>F)
## n                            4 611364  152841 2.394e+05 <2e-16 ***
```

```
## korelacija                   3  75993   25331 3.968e+04 <2e-16 ***
## model                        5   6044    1209 1.893e+03 <2e-16 ***
## alpha                        2  46640   23320 3.653e+04 <2e-16 ***
## n:korelacija                12  74496    6208 9.724e+03 <2e-16 ***
## n:model                     20  11316     566 8.862e+02 <2e-16 ***
## korelacija:model            15   3421     228 3.573e+02 <2e-16 ***
## n:alpha                      8  47235    5904 9.248e+03 <2e-16 ***
## korelacija:alpha             6   8489    1415 2.216e+03 <2e-16 ***
## model:alpha                 10   1122     112 1.757e+02 <2e-16 ***
## n:korelacija:model          60   4922      82 1.285e+02 <2e-16 ***
## n:korelacija:alpha          24   8708     363 5.683e+02 <2e-16 ***
## n:model:alpha               40   1891      47 7.403e+01 <2e-16 ***
## korelacija:model:alpha      30    116       4 6.049e+00 <2e-16 ***
## n:korelacija:model:alpha   120    243       2 3.168e+00 <2e-16 ***
## Residuals                719640 459454       1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```