Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2021-01-05

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. Za izračun intervalov zaupanja bomo uporabili različne metode, ki smo jih spoznali v poglavjih simulacij in metod samovzorčenja. Tako bomo primerjali pokritosti in širine različno ocenjenih intervalov zaupanja. Poleg tega bomo preverjali, če probleme kršenja predpostavk lahko (vsaj delno) odpravimo z uporabo posplošenih 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. V zadnjem delu tega poglavja so predstavljeni načini za izračun intervalov zaupanja regresijskih koeficientov.

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).

Ocenjevanje intervalov zaupanja

Intervale zaupanja bomo ocenili na tri različne načine - naivni, obrnjeni in na podlagi stanardnih napak. Funkciji *lm* in *glm* avtomatično vračata intervale zaupanja na podlagi standardnih napak. Naslednji opisi različnih intervalov zaupanja so povzeti po predavanjih pri predmetu Računsko zahtevne metode.

Naivni

Predpostavljamo, da je porazdelitev razlike med oceno parametra in parametrom simetrična okoli 0. Njihova prednost je, da vedno dajejo vrednosti, ki so možne vrednosti parametra (npr. pri deležu bodo vedno med 0 in 1). Izračunamo jih po formuli: $P(\Theta \in (L, U)) = 1 - \alpha$.

Na podlagi standardnih napak

Najprej izračunamo standardno napako vzorčne ocene ($se(g)$) kot standardni odklon vzorčnih ocen iz bootstrap vzorcev. Potem izračunamo intervale zaupanja po običajih formulah (npr. za normalno porazdelitev $P(g - z_{\alpha/2}se(g) < \theta < g + z_{\alpha/2}se(g)) = 1 - \alpha$. Tako izračunanemu intervalu lahko še odštejemo ocenjeno pristranskost.

Obrnjeni

Tu ločimo dva primera - za lokacijske parametre in parametre merila. Pri lokacijskem parametru predpostavljamo, da je parameter, za katerega ocenjujemo interval zaupanja, paramater "lokacije". Če vsem vrednostim spremenljivke prištejemo a, je nova vrednost parametra $\theta + a$. Slabost tega načina je v tem, da če predpostavka ni izpolnjena, lahko dobimo ne samo nepravilne, ampak tudi nesmiselne intervale (npr. za delež take, ki niso med 0 in 1). Izračunamo ga po sledeči formuli: $P(\theta \in (2\hat{\theta} - U, 2\hat{\theta} - L)) = 1 - \alpha$.

Pri parametru merila pa predpostavljamo, da je parameter, za katerega ocenjujemo interval zaupanja, paramater "merila". Če vse vrednostim spremenljivke pomnožimo z a, je nova vrednost parametra $g(a)\theta$, kjer je funkcija g odvisna le od tipa parametra (varianca,...). Slabost je v tem, da če predpostavka ni izpolnjena, lahko dobimo nepravilne intervale. Izračunamo pa jih po formuli $P(\theta \in (\hat{\theta}^2/U, \hat{\theta}^2/L)) = 1 - \alpha$.

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 ($Gamma(\alpha, \beta)$), kjer bomo parameter α spreminjali tako, da dobimo različno močno asimetrične porazdelitve ($(\alpha, \beta) \in \{(1, 5), (3, 5), (5, 5)\}$;
- 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()* in si s tem olajšamo delo pri generiranju podatkov.

Funkciji lm() in glm()

Funkcija *lm()* se uporablja za generiranje 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 gamma 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).

Pričakovanja

Pri večji korelaciji med pojasnjevalnimi spremenljivkami pričakujemo širše intervale zaupanja regresijskih koeficientov ter večjo pristranost ne glede na izbiro metode.

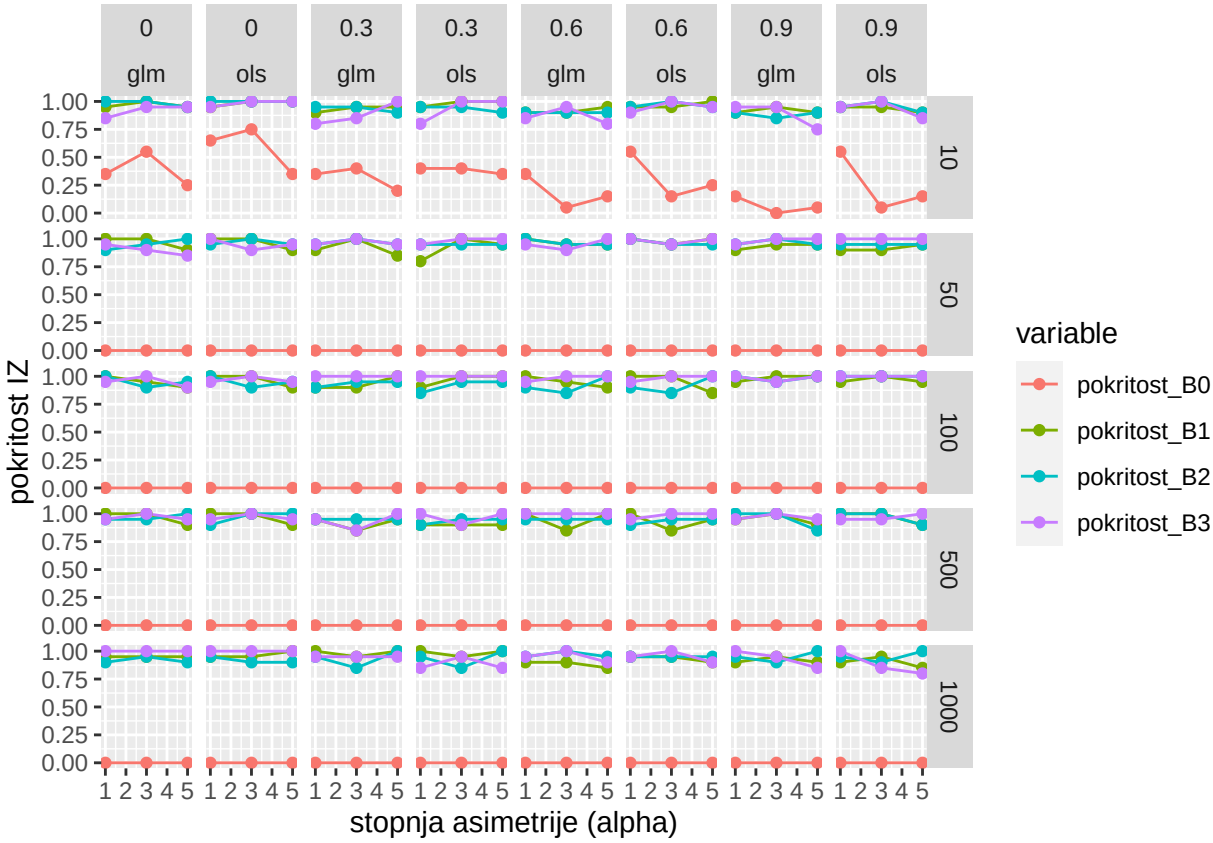
Večje razlike med metodami pričakujemo predvsem pri manjših velikostih vzorcev in večji asimetriji porazdelitve ostankov. Pri dovolj velikih vzorcih pričakujemo podobne rezultate obeh metod, prav tako pa seveda tudi manjšo variabilnost rezultatov.

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()*.

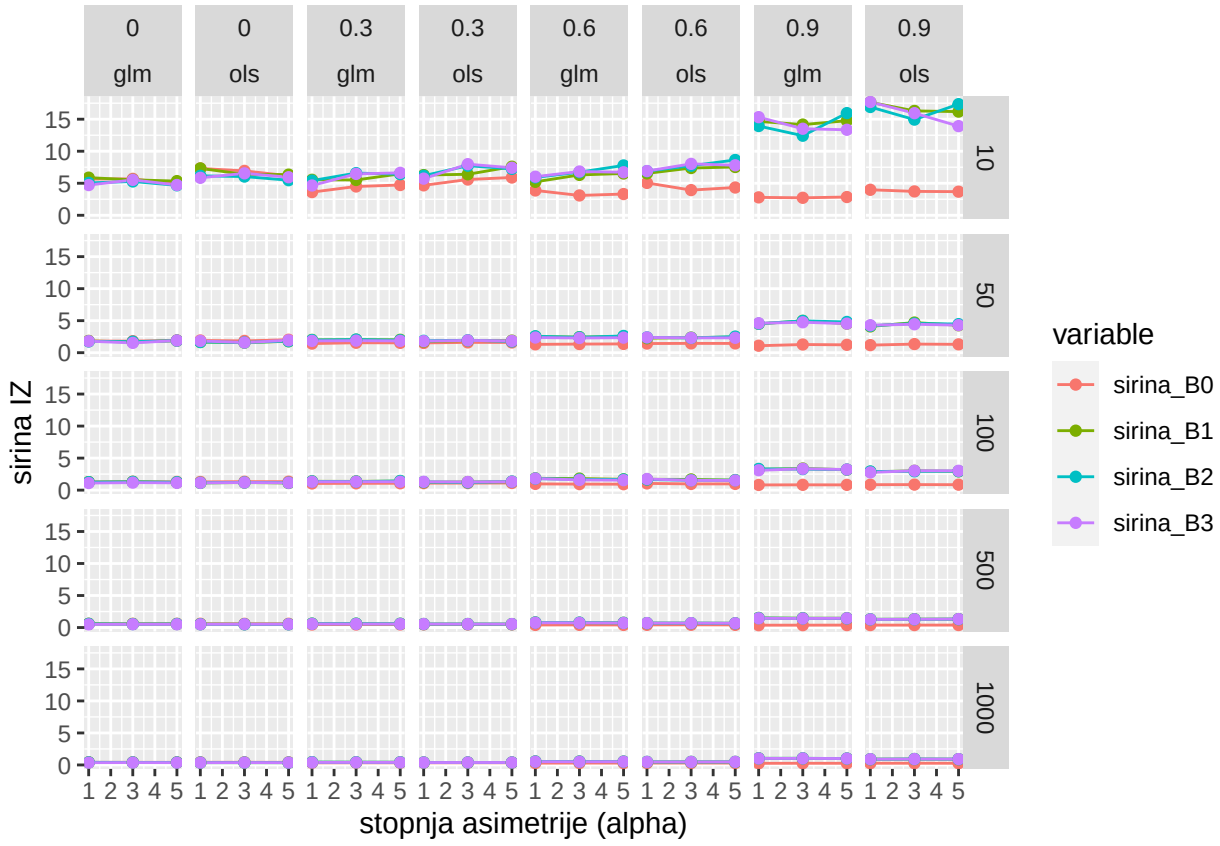
V primeru, ko iz modela izločimo spremenljivko X_3, ne pričakujemo posebnih sprememb v rezultatih, saj spremenljivka nima vpliva na vrednost pojasnjevalne spremenljivke. V primeru, ko izločimo spremenljivko X_2, pa pričakujemo spremembe v rezultatih - širše intervale zaupanja regresijskih koeficientov 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 porazelitve prišlo do manjše pokritosti in večje širine intervalov zaupanja.

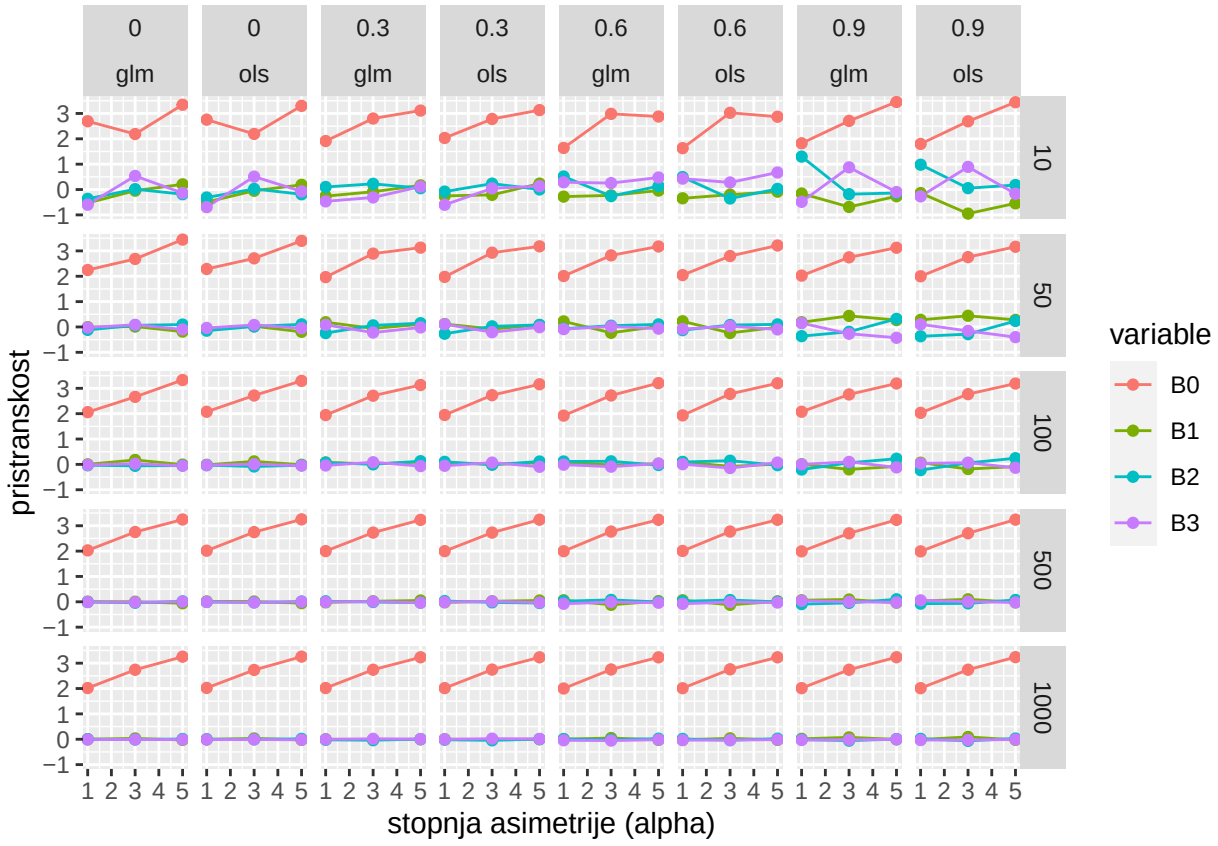
Predstavitev rezultatov



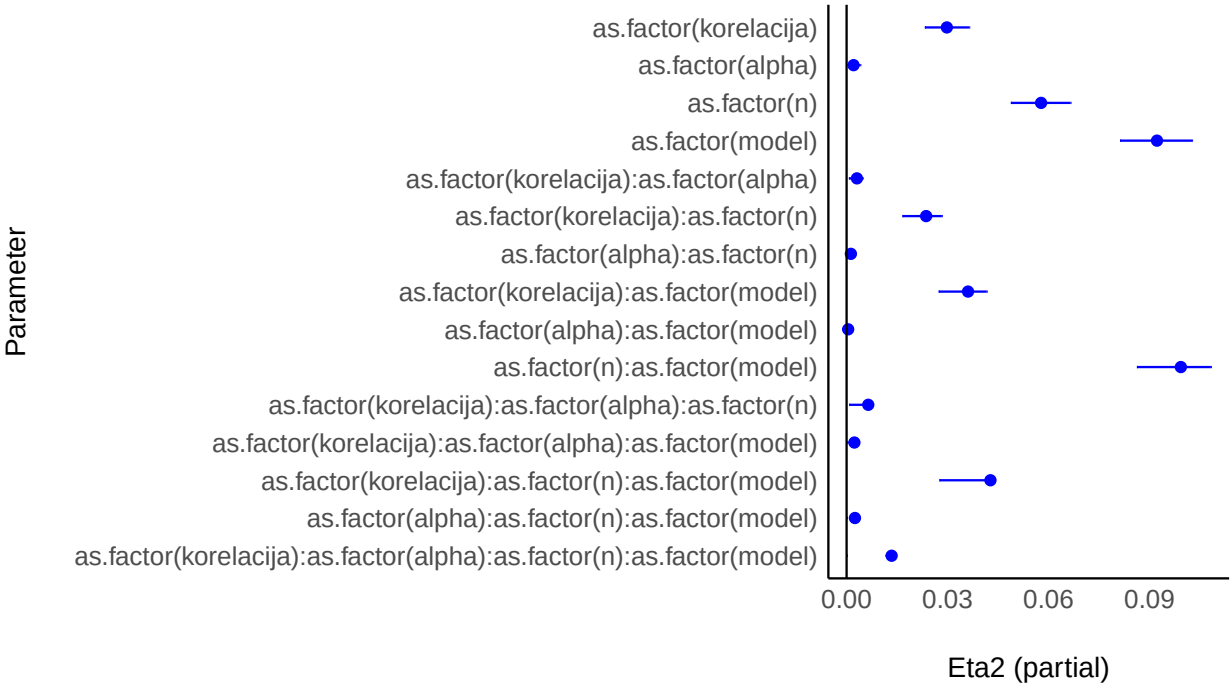
Slika 1: Pokritost intervalov zaupanja

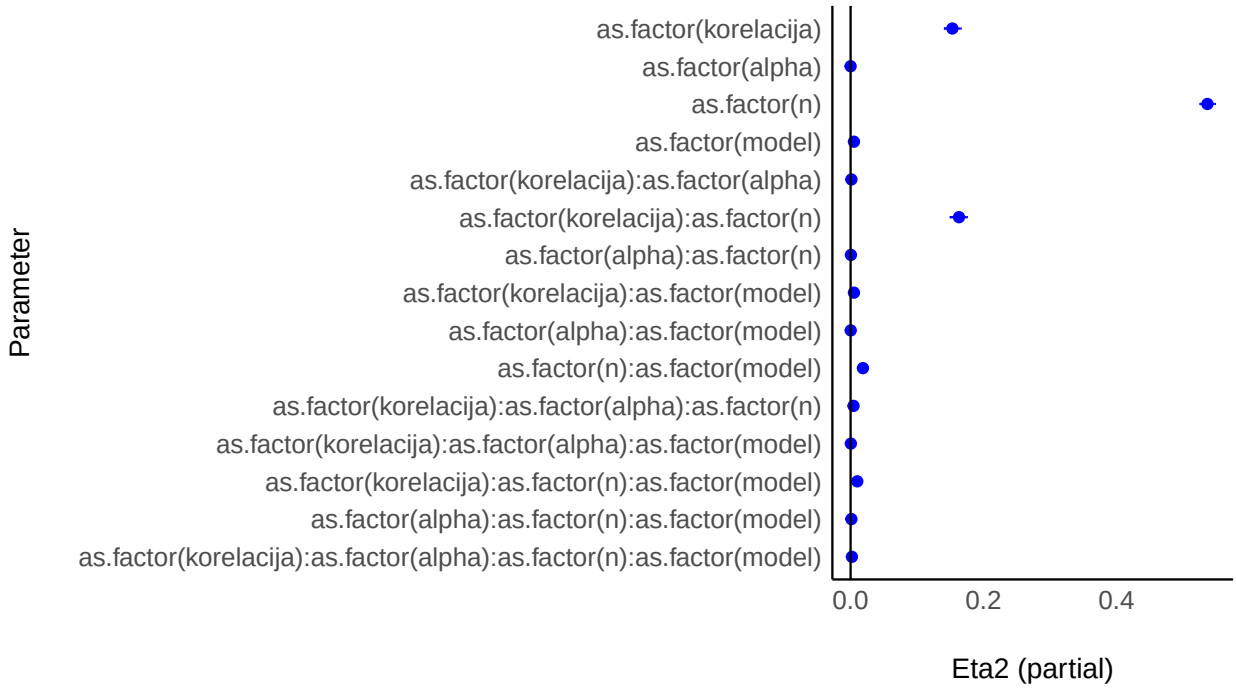


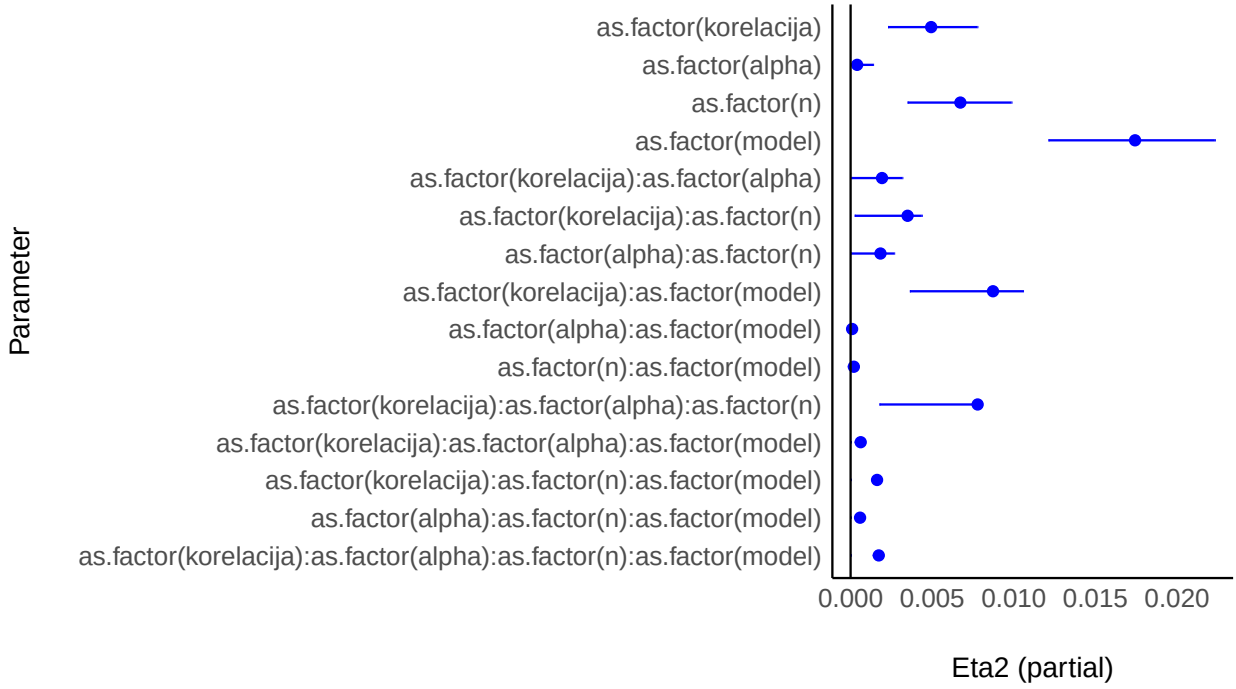
Slika 2: Širina intervalov zaupanja



Slika 3: Pristranskost ocen







Ugotovitve

Nek povzetek, zaključek.

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.
- 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.
- https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html

Priloge

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