

Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2021-01-01

Kazalo vsebine

Uvod	2
Teoretični del	2
Posplošeni linearni modeli	2
Linearna regresija	2
Metoda najmanjših kvadratov (MNK)	4
Metoda največjega verjetja (MNV)	4
Restrignirana metoda največjega verjetja (RMNV)	6
Funkcija glm()	6
Bootstrap?	6
Ocenjevanje intervalov zaupanja	6
Generiranje podatkov	7
Predstavitev rezultatov	7
Ugotovitve	7
Viri	7
Priloge	8

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. 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 modelov.

Treba je še zapisat, kakšne rezultate pričakujeva. Najbolje, da kar tu narediva še en odstavek. Razmisliti je potrebno tudi, kako bi se dalo rezultate izboljšati.

Teoretični del

Tu predstaviva metode, ki jih uporabljava ali primerjava. Poudarek je na predpostavkah in ostalih značilnostih, ki jih preverjava.

Naloga je osredotočena na linearno regresijo, ki spada pod posplošene linearne modele. Najprej so 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. Uporabni so na mnogo različnih področjih (npr. v medicini, financah, izobraževanju, ...). Za modeliranje pridejo prav, kadar so meritve na testirancih opravljene večkrat skozi neko časovno obdobje (Maver, 2018, str. 6).

Linearna regresija

Linearna regresija je statistični model, ki ga v osnovni 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:

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

V tem modelu je predpostavljeno, da so regresijski koeficienti fiksni, obstajajo pa primeri, v katerih je smiselno predvidevati, da so nekateri izmed koeficientov naključni (Jiang, 2007, str. 1).

Ključne predpostavke linearnega modela (Correlation and regression with R, 2016) so naslednje:

1.) Normalna porazdelitev: za katerokoli vrednost X je Y porazdeljena normalno.

Pri tej predpostavki ponavadi navajamo normalno porazdeljenost ostankov. Če velja $\epsilon \sim N(0, \sigma)$, iz tega sledi $Y \sim N(\cdot, \sigma)$

2.) Homogenost variance (homoskedastičnost): varianca residualov je enaka za vse vrednosti X ;

3.) Linearna zveza med pojasnjevalno spremenljivko X in povprečjem odzivne spremenljivke Y ;

4.) Opazovanja so med seboj neodvisna.

5.) 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.

6.) Medsebojna neodvisnost vrednosti slučajne spremenljivke ϵ prav tako spada med temeljne predpostavke linearnega regresijskega modela. Pravimo, da spremenljivke niso avtokorelirane oz. da v modelu ni avtokorelacije. To predpostavko zapišemo $Cov(\epsilon_i, \epsilon_j) = 0$ za vsak $i \neq j$ (Pfajfar, 2018).

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: $cov(e_i, e_j | x_i, x_j) = 0; i \neq j$
- 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: $u_i \sim N(0, \sigma_i^2)$. Posledično je odvisna spremenljivka y tudi normalno porazdeljena s.s.: $y_i \sim N(\beta_1 x_{1i} + \dots + \beta_k x_{ki}, \sigma_i^2)$

Prva metoda, ki jo lahko uporabimo pri modeliranju v programskem jeziku *R*, je metoda najmanjših kvadratov (MNK), ki je najbolj enostavna. Nato bom opisala metodo največjega verjetja (MNV) in restringirano metodo največjega verjetja (RMNV), ki je bila razvita zaradi pomanjkljivosti prejšnje. Pri obeh standardnih metodah (MNV in RMNV) ločimo dva načina ocenjevanja in sicer točkovno ocenjevanje (za majhen vzorec) in ocenjevanje z asimptotično kovariančno matriko (za velike vzorce). V delu bom metodi predstavila le za točkovno ocenjevanje. V primeru osnovnega modela linearne regresije sta oceni koeficientov po metodah MNK in MNV ekvivalentni, če predpostavimo normalno porazdelitev ϵ in α .

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

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

Metoda največjega verjetja (MNV)

Naj bodo X_1, \dots, X_n n.e.p. s.s., porazdeljene z gostoto $f_Y(x; \theta_1, \dots, \theta_k)$. Funkcijo

$$L(\theta; x) = L(\theta_1, \dots, \theta_k; x_1, \dots, x_n) = \prod_{i=1}^n f_Y(x_i; \theta_1, \dots, \theta_k)$$

imenujemo funkcija verjetja za vzorec velikosti n .

Za vsak vzorec x naj bo $\hat{\theta}(x)$ vrednost parametra, v katerem funkcija L doseže maksimum. Cenilka po MNV za parameter θ na osnovi vzorca X je $\hat{\theta}(X)$. Ta cenilka je pod ustreznimi predpostavkami dosledna, vendar ne nujno nepristranska.

MNV ima asimptotsko najmanjšo varianco (je učinkovita), kar sledi kot posledica neenakosti Cramér-Rao. V primeru diskretne porazdelitve je verjetje kar produkt verjetnosti:

$$L(x; \theta) = \prod_{i=1}^n P(X_i = x_i; \theta).$$

Gaussov linearni mešani model lahko predstavimo z njegovo marginalno porazdelitvijo. Tako je Y porazdeljen kot

$$Y \sim N(X\beta, V),$$

kjer je $V = R + ZGZ^T$ ter $R = \text{diag}(R_1, \dots, R_n)$, $G = \text{diag}(G_1, \dots, G_n)$ in $Z = \text{diag}(Z_1, \dots, Z_n)$. Vektor $\theta = (R_1, \dots, R_n, G_1, \dots, G_n)$ predstavlja vektor vseh variančnih komponent, vključenih v V . V nadaljevanju je predstavljena izpeljava cenilke po MNV, kjer y in Y označujejo vektorje.

Porazdelitev slučajnega vektorja Y je dana z gostoto

$$f(y) = \frac{1}{(2\pi)^{\frac{n}{2}} |V|^{\frac{1}{2}}} e^{-\frac{1}{2}(y-X\beta)^T V^{-1}(y-X\beta)},$$

kjer je n dimenzija Y . Torej je logaritmirana funkcija verjetja enaka

$$l(\beta, \theta) = c - \frac{1}{2} \log(|V|) - \frac{1}{2} (y - X\beta)^T V^{-1} (y - X\beta),$$

kjer je c neka konstanta. Ko odvajamo po parametru β , dobimo

$$\frac{\partial l}{\partial \beta} = X^T V^{-1} Y - X^T V^{-1} X \beta.$$

Odvod po komponenti vektorja θ pa je enak

$$\frac{\partial l}{\partial \theta_r} = \frac{1}{2} \{ (Y - X\beta)^T V^{-1} \frac{\partial V}{\partial \theta_r} V^{-1} (Y - X\beta) - \text{sled}(V^{-1} \frac{\partial V}{\partial \theta_r}) \}, r = 1, \dots, q,$$

kjer je θ_r r -ta komponenta vektorja $\theta = (R_1, \dots, R_n, G_1, \dots, G_n)$. Le ta je dimenzije q .

Cenilka bo rešitev vektorskih enačb

$$\frac{\partial l}{\partial \beta} = 0, \frac{\partial l}{\partial \theta} = 0.$$

Dobljena cenilka za vektor koeficientov po MNV je

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} Y.$$

Podrobnejšo izpeljavo z vsemi potrebnimi predpostavkami je v svojem delu predstavil Jiang (Jiang, 2007).

Da še malo poenostavimo, pogledjmo primer osnovnega modela linearne regresije v matrični obliki $Y = X\beta + \epsilon$, kjer so

$$\begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,p} \\ X_{2,1} & X_{2,2} & \dots & X_{2,p} \\ \vdots & \vdots & & \vdots \\ X_{n,1} & 0 & \dots & X_{n,p} \end{bmatrix}, \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Tu je matrika X dimenzije $n \times (p+1)$ in p število spremenljivk. Predpostavimo, da je matrika X polnega ranga in velja

$$\epsilon \sim N(0, \sigma^2 I_{n \times n}).$$

Cenilka, dobljena po MNV, je

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Ta cenilka je nepristranska, saj velja

$$E(\hat{\beta}) = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X \beta = \beta.$$

Izpeljava variančno-kovariančne matrike:

$$\begin{aligned} \text{var}(\hat{\beta}) &= \text{var}((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T \text{var}(Y) X (X^T X)^{-1} = \\ &= (X^T X)^{-1} X^T \sigma^2 I_{n \times n} X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}. \end{aligned}$$

Velja

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 (X^T X)^{-1}_{ii}}} \sim N(0, 1).$$

\

V primeru, ko je $p = 1$, je ocena regresijskega koeficienta β_1 po metodah MVN in MNK enaka

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Restriringirana metoda največjega verjetja (RMNV)

Zaradi pomanjkljivosti metode največjega verjetja se je razvila restringirana metoda največjega verjetja. Ta ima več prednosti pred MNV, med drugim omogoča pridobivanje cenilk zgolj za parametre, ki so predmet zanimanja in se ne ozira na nezanimive parametre. Cenilke, dobljene po RMNV, so invariantne za fiksne učinke modela, metoda pa le-teh niti ne ocenjuje (Maver, 2018, str.17).

Poglejmo še podrobnejši opis RMNV, povzet po obsežnejši literaturi (Jiang, 2007). Naj bo $Y \sim N(X\beta, V)$. Brez škode za splošnost naj bo izpolnjena predpostavka, da ima matrika $X \in R^{m \times p}$ poln rang, torej $\text{rang}(X) = p$. Naj bo A taka matrika dimenzije $m \times (m-p)$, za katero velja $\text{rang}(A) = m-p$ in $A^T X = 0$. Za potrebe izpeljave naj bo $Z = A^T Y$. Velja: $E(A^T Y) = A^T X\beta = 0$, $\text{Var}(A^T Y) = A^T V A \Rightarrow Z \sim N(0, A^T V A)$. Sledi, da je zvezna porazdelitvena gostota slučajne spremenljivke Z dana kot

$$f_{Z,R}(z) = \frac{1}{(2\pi)^{(m-p)/2} |A^T V A|^{1/2}} e^{-\frac{1}{2} z^T (A^T V A)^{-1} z}.$$

Indeks R tu označuje, da gre za restringirano metodo. Restriringirana logaritmirana funkcija verjetja je

$$l_R(\theta) = c - \frac{1}{2} \log(|A^T V A|) - \frac{1}{2} Z^T (A^T V A)^{-1} Z, \quad (1)$$

kjer je c konstanta. Z odvajanjem pa dobimo

$$\frac{\partial l_R}{\partial \theta_i} = \frac{1}{2} \{Y^T P \frac{\partial V}{\partial \theta_i} P Y - \text{sled}(P \frac{\partial V}{\partial \theta_i})\}, i = 1, \dots, q.$$

Tu je $P = A(A^T V A)^{-1} A^T$. Cenilka po restringirani metodi največjega verjetja je definirana kot točka, v kateri funkcija (1) zavzame maksimum. Dobi se jo kot rešitev enačbe $\frac{\partial l_R}{\partial \theta} = 0$ (Jiang, 2007, str. 13). RMNV je torej metoda, s katero se direktno ne pridobi cenilke parametra β , pač pa le cenilko parametra θ . Razlog je v tem, da se β izloči že pred ocenjevanjem. Cenilko za β se dobi iz enačbe $V = V(\hat{\theta})$, kjer je $\hat{\theta}$ cenilka za θ , dobljena z RMNV. Izpeljave in podrobnejšo razlago metode lahko pogledamo v (Jiang, 2007).

Funkcija glm()

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.

Bootstrap?

Ocenjevanje intervalov zaupanja

N načini, kako bova ocenjevali IZ. Lahko še prilagodiva, zaenkrat pa predlagam:

- naivni
- na podlagi standardnih napak
- obrnjeni

Na prosojnicah so zapisane glavne značilnosti posameznih IZ. Najbrž je dovolj, če zapiševa zelo na kratko

Generiranje podatkov

Natančen opis generiranja podatkov.

Fiksni parametri pri generiranju podatkov so sledeči:

- porazdelitev pojasnjevalnih spremenljivk:
 - $X_1 \sim \text{Gamma}(2, 5)$
 - $X_2 \sim \text{Gamma}(2, 5)$
 - $X_3 \sim \text{Gamma}(5, 5)$
 - $X_4 \sim \text{Gamma}(5, 5)$
 - $X_5 \sim \text{Gamma}(5, 5)$
- formula za generiranje podatkov:

$$y_i = 5x_1 + x_2 + 5x_3 + x_4 + 0x_5 + \epsilon_i$$

Pri generiranju podatkov bomo spreminjali sledeče:

- velikost vzorca $n \in \{10, 20, 30, 50, 100, 500, 1000\}$
- korelacija med odvisnimi spremenljivkami ($cor \in \{0, 0.3, 0.6, 0.9\}$)
- porazdelitev napak ($\text{Gamma}(\alpha, \beta)$), kjer bomo parameter α spreminjali tako, da dobimo različno močno asimetrične porazdelitve ($(\alpha, \beta) \in \{(1, 5), (2, 5), (2, 2), (5, 5)\}$)
- v modelu ne upoštevamo vseh neodvisnih spremenljivk (spreminjamo število spremenljivk, ki jih upoštevamo)

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

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

Predstavitev rezultatov

Predstavitev rezultatov (samo grafično ni dovolj, potrebno je še analizirati varianco na rezultatih).

Ugotovitve

Viri

Sproti navajaj vsaj linke, da potem na koncu samo uredi in jih pravilno navede

- 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.

- D. Bates et al., *Package 'lme4'*, v: Linear Mixed-Effects Models using 'Eigen' and S4, [ogled 4.6.2019], dostopno na <https://cran.r-project.org/web/packages/lme4/lme4.pdf>.
- *lme*, v: RDocumentation, [ogled 15.5.2019], dostopno na <https://www.rdocumentation.org/packages/nlme/versions/3.1-137/topics/lme>.
- *glm*, v: RDocumentation, [ogled 15.5.2019], dostopno na <https://www.rdocumentation.org/packages/stats/versions/3.6.0/topics/glm>.
- M. Raič, *O linearni regresiji*, 2014. Najdeno na spletnem naslovu: Linearna regresija
- L. Pfajfar, *Osnovna ekonometrija*, učbeniki Ekonomske fakultete, Ljubljana, 2018.
- Correlation and regression with R

Priloge

Rmd datoteka s kodo, ali pa če dava kar povezavo na github repozitorij