Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2020-12-31

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 modelov.

Pričakovanja

Nekam je treba zapisat, kakšne rezultate pričakujeva, ni pa nujno, da je to svoje poglavje. Razmislit je potrebno tudi, kako bi se dalo rezultate izboljšati.

Obravnavane metode

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

Linearna regresija

Linearna regresija je statistični model, ki ga lahko zapišemo v naslednji obliki:

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

, kjer je $\beta \in \mathbb{R}^k$ vektor regresijsih koeficientov, $X \in \mathbb{R}^{k \times n}$ matrika pojasnjevalnih spremenljivk in $Y \in \mathbb{R}^n$ vektor odvisnih spremenljivk.

Najpomembnejše in najpogosteje navedene predpostavke linearnega modela:

- Ostanki so porazdeljeni normalno: $\epsilon \sim N(0, \sigma)$
- Homogenost variance

Verbič 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: $Cov(x_2, u) = Cov(x_3, u) = ... = Cov(x_k, u) = 0$
- 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: $u_i \sim N(0, \sigma_i^2)$. Posledično je odvisna spremenljivka y tudi normalno porazdeljena s.s.: $y_i \sim N(\beta_1 x_{1i} + ... + \beta_k x_{ki}, \sigma_i^2)$

Minimiziramo vsoto kvadratov residualov: $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

Posplošeni linearni modeli

Treba je raziskat funkcijo `glm()`. Naj bi bila fajn, ker lahko za napake nastavimo kakšno drugo porazdelitev (tako je rekel profesor).

Bootstrap?

Ocenjevanje intervalov zaupanja

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

- naivni
- na podlagi standardnih napak
- obrnjeni

Generiranje podatkov

Natančen opis generiranja podatkov.

Fiksni parametri pri generiranju podatkov so sledeči:

- porazdelitev pojasnjevalnih spremenljivk:
  - $X_1 \sim Gamma(2, 5)$
  - $X_2 \sim Gamma(2, 5)$
  - $X_3 \sim Gamma(5, 5)$
  - $X_4 \sim Gamma(5, 5)$
  - $X_5 \sim Gamma(5, 5)$
- 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 ($Gamma(\alpha, \beta)$), 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 $X_i \sim Gamma(k_i, \theta)$, potem je

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

Predstavitev rezultatov

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

Ugotovitve

Viri

- M. Raič, *O linearni regresijji*, 2014. Najdeno na spletnem naslovu: Linearna regresija
- M. Verbič, L. Pfajfar, R. Rogelj, *Ekonometrični obrazci in postopki: Dopolnjena druga izdaja.* Ljubljana: Ekonomska fakulteta, 2017. [ISBN 9789612403157]

Priloge

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