

Vpliv kršenja predpostavk linearne regresije na njene rezultate

Seminarska naloga pri predmetu Računsko zahtevne metode

Anja Žavbi Kunaver in Vesna Zupanc

2020-12-30

Kazalo vsebine

Uvod	2
Pričakovanja	2
Obravnavane metode	2
Linearna regresija	2
Posplošeni linearni modeli	2
Bootstrap?	2
Ocenjevanje intervalov zaupanja	2
Generiranje podatkov	3
Predstavitev rezultatov	3
Ugotovitve	3
Viri	3
Priloge	3

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.

Pričakovanja

Nekam je treba zapisat, kakšne rezultate pričakujeva, ni pa nujno, da je to svoje poglavje. Razmisliti 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 regresijskih koeficientov, $X \in \mathbb{R}^{k \times n}$ matrika pojasnjevalnih spremenljivk in $Y \in \mathbb{R}^n$ vektor odvisnih spremenljivk.

Predpostavke linearnega modela:

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

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

POGLEJ V PREDMET EKONOMETRIJA!

Posplošeni linearni modeli

Treba je raziskati 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:
 - $X_1 \sim \text{Gamma}(2, 5)$
 - $X_2 \sim \text{Gamma}(2, 5)$
 - $X_3 \sim \text{Gamma}(5, 5)$
 - $X_4 \sim \text{Gamma}(5, 5)$
 - $X_5 \sim \text{Gamma}(5, 5)$
- 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

- M. Raič, *O linearni regresiji*, 2014. Najdeno na spletnem naslovu: Linearna regresija

Priloge

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