Robustifikace hřebenové regrese - první pokusy

Tomáš Jurczyk

Matematicko-fyzikální fakulta, UNIVERZITA KARLOVA V PRAZE, Katedra pravděpodobnosti a matematické statistiky, Sokolovská 83, 18675 Praha 8, ČR

1 Abstrakt

Jeden z problémů, se kterými se můžeme při datové analýze setkat, je multikolinearita (téměř lineární závislost mezi regresory). Multikolinearita vede k nestabilitě řešení normální rovnice a může způsobit velký rozptyl odhadu regresních koeficientů. Jednou z metod, která se snaží tento problém řešit je hřebenová regrese. Tato metoda ale není imunní vůči přítomnosti kontaminace, která může multikolinearitu skrýt nebo ji naopak uměle vytvořit. Tento příspěvek se věnuje prvním jednoduchým pokusům o robustifikaci metody hřebenové regrese. V práci jsou představeny dva návrhy založené na robustní metodě nejmenších vážených čtverců (LWS).

2 Multikolinearita

Značení

Předpokládejme lineární regresní model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, kde regresní matice má plnou hodnost $h(\boldsymbol{X}) = k$, $\boldsymbol{e} \sim (\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.
$\boldsymbol{b}$ je klasický odhad $\boldsymbol{\beta}$ metodou nejmenších čtverců.

Co je to multikolinearita

Termín multikolinearita nebo také kolinerita označuje situaci, kdy jsou regresory „téměř" lineárně závislé.

Důsledky

Nechť $\boldsymbol{X} = \boldsymbol{P}\boldsymbol{T}\boldsymbol{Q}'$ je rozklad matice $\boldsymbol{X}$ podle singulárních hodnot.

- $\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I_k}$
- $\boldsymbol{T}$ diagonální s $t_1 \geq t_2 \geq \ldots \geq t_k > 0$
- Za multikolinearity je hodnota t_k, případně i dalších t_j, velmi malá.

Dá se ukázat, že:

$$E\,\|\boldsymbol{b}\|^2 = \|\boldsymbol{\beta}\|^2 + \sigma^2 \sum_{i=1}^{k} t_i^{-2}$$

$$\mathrm{var}\boldsymbol{b} = \sigma^2 \sum_{i=1}^{k} t_i^{-2} \boldsymbol{q}_i \boldsymbol{q}_i'$$

$\boldsymbol{q}_i$ sloupce matice $\boldsymbol{Q}$.

Multikolinearita způsobuje velkou očekávanou hodnotu délky vektoru $\boldsymbol{b}$ a může způsobit velký rozptyl statistiky b_j.

Detekce

Jednou z možností zjišťování multikolinearity je využití indexů podmíněnosti:

$$\eta_j = t_1/t_j, j = 1, \ldots, k.$$

Tyto je výhodné takto definovat především pro $\boldsymbol{X}$ s normovanými sloupci.

Použití – v literatuře se uvádí toto:

$\eta_j < 10(30)$	– vše v pořádku
$\eta_j \in (10(30), 100)$	– podezření na multikolinearitu, je nutná opatrnost
$\eta_j > 100$	– jedná se o silný vztah mezi regresory

Výsledky převzaty z knihy [Zvára (1989)].

3 Hřebenová regrese

Jak řešit problém s multikolinearitou?

Místo klasického odhadu se v případě podezření na multikolinearitu raději použije vychýlený odhad:

Definice hřebenové regrese

Pro $\delta \geq 0$ definujeme hřebenový odhad vektoru $\boldsymbol{\beta}$ jako

$$\boldsymbol{b}_\delta = (\boldsymbol{X}'\boldsymbol{X} + \delta \boldsymbol{I})^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

- ▶ Tento odhad má při malých hodnotách δ (přesněji $0 < \delta < 2\sigma^2 \|\boldsymbol{\beta}\|^{-2}$) menší čtvercovou chybu než klasický odhad $\boldsymbol{b}$.

Volba δ

Možnosti, jak volit δ:

a) Odhadne se horní mez (viz předešlá vlastnost): $\delta = 2s^2 \|\boldsymbol{b}\|^{-2}$, kde s^2 je odhad σ^2. Někdy se také doporučuje $\delta = ks^2 \|\boldsymbol{b}\|^{-2}$.

b) Graficko metodu zvanou ridge trace, kde se vynášejí do grafu hřebenové odhady koeficientů proti proměnné δ. Za vhodnou volbu δ se pak považuje podle autorů metody taková hodnota, při níž jsou již znaménka a hodnoty složek $\boldsymbol{b}_\delta$ stabilizovány.

Problémy

Hřebenová regrese není díky své konstrukci uzpůsobena pro práci s odlehlými pozorováními. Navíc odlehlá pozorování mohou multikolinearitu skrýt nebo ji naopak uměle vytvořit – už i jedno přidané odlehlé pozorování může výrazně změnit, jak výběrový korelační koeficient, tak například i výše zmíněné indexy podmíněnosti.

Detekce multikolinearity v kontaminovaných datech je velmi důležitá, protože odhaluje pravé závislosti v datech.

Reference

[Rousseeuw and Leroy (1987)] Rousseeuw, P. J., and Leroy, A. M., Robust Regression and Outlier Detection, *John Wiley & Sons*, 1987.

[Víšek (2001)] Víšek, J. Á., Regression with high breakdown point, *Robust 2000 (eds. Jaromír Antoch & Gejza Dohnal, published by Union of Czech Mathematicians and Physicists), Prague: matfyzpress*, 324-356, 2001.

[Víšek (2008)] Víšek, J. Á., Consistency of the Instrumental Weighted Variables, To appear in *Annals of the Institute of Statistical Mathematics*, 2008.

[Zvára (1989)] Zvára, K., Regresní analýza, *Academia, Praha*, 1989.

4 Nejmenší vážené čtverce

Nejprve definujme metodu vážených nejmenších čtverců:

Definice WLS

Nechť $w_i > 0, i = 1 \ldots, n$ jsou váhy. Potom odhad

$$\boldsymbol{b}^{(WLS,w)} = \underset{\beta \in \mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i r_i^2(\boldsymbol{\beta})$$

nazveme odhadem metodou vážených nejmenších čtverců (WLS – the weighted least squares).
Ekvivalentně

$$\boldsymbol{b}^{(WLS,w)} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y},$$

kde n je počet pozorování, r_i je reziduum pro i-té pozorování a $\boldsymbol{W} = diag(w_1, \ldots, w_n)$.

- ▶ Takto definované vážené nejmenší čtverce jsou vlastně klasickou regresí na data $\mathbf{W}^{\frac{1}{2}}\mathbf{y}$, $\mathbf{W}^{\frac{1}{2}}\mathbf{X}$. Každé pozorování je převáženo.

Nyní definujme metodu, která byla představena v článku [*Víšek (2001)*]. Jedná se o robustní metodu, kterou bychom rádi využili při robustifikaci hřebenové regrese.

Definice LWS

Nechť $1 = w_1 \geq w_2 \geq \ldots . \geq w_n \geq 0$ jsou váhy. Potom

$$\boldsymbol{b}^{(LWS,w)} = \underset{\beta \in \mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i r_{(i)}^2(\boldsymbol{\beta})$$

se nazývá odhad metodou nejmenších vážených čtverců (LWS – the least weighted squares). $r_{(n)}^2 \geq \ldots \geq r_{(1)}^2$ jsou uspořádané kvadráty reziduí.

- ▶ Všimněme si, že jednotlivé váhy nejsou přímo přiřazeny konkrétnímu pozorování, metoda sama přiřadí váhy implicitně k pozorováním.
- ▶ Výsledný odhad LWS je vlastně odhad WLS s váhovou maticí $\boldsymbol{W}(\pi)$, která vznikne jistou permutací prvků na diagonále matice $\boldsymbol{W}$, tedy
$$\boldsymbol{b}^{(LWS,w)} = (\boldsymbol{X}'\boldsymbol{W}(\pi)\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{W}(\pi)\boldsymbol{y},$$
kde π je permutace zajišťující minimalizaci kriteriální funkce LWS.

Algoritmus výpočtu LWS[1]

Dva cykly – vnitřní cyklus vybírá metodou iterativně vážených nejmenších čtverců kandidáta na dosažení minima (je potřeba iterační metoda, protože přiřazení vah závisí na reziduích a rezidua závisí na odhadu koeficientů). Vnější cyklus zaznamenává do té doby nejlepší model a také kolikrát byl dosažen. Pokud po dosažení nejlepšího modelu dosaženo opakovaně (řekněme 20 krát) algoritmus končí a tento model je prohlášen za odhad LWS.

[1] *Podrobější informace o algoritmu například v článku [Víšek (2008)].*

5 Robustifikace hřebenové regrese

V této části budou představeny dva první jednoduché návrhy na robustifikaci metody hřebenové regrese za použití nejmenších vážených čtverců. Také bude shrnuto, jak si tyto metody vedly v případech, kdy byla multikolinearita kontaminací zakrytá.

Hřebenová regrese a WLS

Nejdříve je potřeba si uvědomit, že volba vah u WLS neřeší problém se závislostí mezi regresory. Jednotlivá pozorování jsou převážena, to ale nepotlačí závislosti mezi jednotlivými složkami pozorování. Proto pokud je problém s multikolinearitou u regresní matice $\boldsymbol{X}$, bude tento problém i u $\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}$. Hřebenovou regresi pro WLS má tedy smysl definovat jako $\boldsymbol{b}_\delta^{WLS,w} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}.$

Hřebenová regrese a LWS (robustní hřebenová regrese)

Návrh 1

Popis metody

Dvoukrokový odhad:

1.krok: Nejdříve se spočítá LWS

2.krok: V prvním kroku získané $\boldsymbol{W}(\pi)$ využijeme pro výpočet klasické hřebenové regrese na data $\boldsymbol{W}(\pi)^{\frac{1}{2}}\boldsymbol{y}$, $\boldsymbol{W}(\pi)^{\frac{1}{2}}\boldsymbol{X}$. Tedy $\boldsymbol{b}_\delta^{LWS,w} = (\boldsymbol{X}'\boldsymbol{W}(\pi)\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{W}(\pi)\boldsymbol{y}$.

Idea aneb proč by to mohlo fungovat

Máme data, kde odlehlá pozorování zakrývají problém s multikolinearitou.
Indexy podmíněnosti toto neodhalí, také ridge trace vypadá odlišně než u případu s téměř závislými sloupci.
Pokud použijeme LWS s váhami, kde $w_i = 0$ pro $i > h$ pro nějaké h a tato metoda odhalí všechna odlehlá pozorování, pak na datech $\boldsymbol{W}(\pi)^{\frac{1}{2}}\boldsymbol{y}$, $\boldsymbol{W}(\pi)^{\frac{1}{2}}\boldsymbol{X}$ bude již kolinearita patrná a na tato nová data se pak použije hřebenová regrese.

Poznámka: Pokud by toto fungovalo, pak by se v takových případech dala LWS výhodně použít jako robustní detektor multikolinearity.

Výsledky

Ačkoli se zdají být tyto úvahy správné, na simulovaných a také na datech z knížky [*Zvára (1989)*] (kapitola 9.) se ukázalo, že takto navržená metoda nefunfuje, jak bychom čekali.
Za přítomnosti odlehlých pozorování zakrývajících multikolinearitu LWS neodhalí správně tato odlehlá pozorování, ačkoli je na to „ stavěná". Navíc dává přiřazením vah vždy alespoň jednomu z odlehlých velmi vysokou důležitost.

Interpretace tohoto jevu

Jsem přesvědčen, že tento problém je způsoben tím, že v datech je vlastně jakási volnost způsobená téměř závislostí. V datech jsou v zásadě dvě skupiny regresorů, které se snaží vysvětlit to samé.
Při přidání odlehlého pozorování k datům se potom může jedna z těchto skupin soustředit výhradně na vyrovnání jen tohoto pozorování – přitom může být toto vyrovnání velmi přesné. Což vede k tomu, že LWS vybere raději tento model s přesně vyrovnaným odlehlým pozorováním místo modelu na skutečných datech, která jsou ovšem nestabilní.

Závěr

- ▶ Vlivem skryté multikolinearity nedokáže navržená metoda ve svém prvním kroku úspěšně odhalit všechna odlehlá pozorování, a proto není vhodným kandidátem pro robustní verzi hřebenové regrese.
- ▶ V knize [*Rousseeuw and Leroy (1987)*] doporučují autoři řešit možný problém skryté multikolinearity tím, že také nejdříve očistí data od odlehlých pozorování metodou nejmenšího mediánu čtverců a poté provedou hřebenovou regresi.
Vzhledem k předešlým úvahám a provedeným analýzám se obávám, že tato metoda také nemůže fungovat správně.

Návrh 2

Popis metody

Pro pevné δ se do vnitřního cyklu LWS implementuje iteračně

$$\boldsymbol{b}_{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}(\pi_{(k)})\boldsymbol{X} + \delta\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{W}(\pi_{(k)})\boldsymbol{y}$$

místo

$$\boldsymbol{b}_{(k+1)} = (\boldsymbol{X}'\boldsymbol{W}(\pi_{(k)})\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}(\pi_{(k)})\boldsymbol{y}.$$

Vnější cyklus pak vyhodnocuje nejlepší model podle klasického kritéria LWS.

Nevýhody

Pro zkonstruování ridge trace se provede takto upravená metoda pro nějakou mříž hodnot δ. Nic ale nezaručuje, že s rostoucím δ zůstane výsledná ideální permutace vah stále stejná, což potvrdily i provedené výpočty. V ridge trace se tedy mohou vyskytovat skoky.
Tím se komplikuje výběr vhodného δ. Navíc je také otázka, co vůbec znamenají výsledky a) na volbu δ v části o hřebenové regresi, protože odhad z celých dat za přítomnosti kontaminace může být touto kontaminací výrazně ovlivněn, nehledě na další problémy s tím spojené.

Výhody

S rostoucím δ se narozdíl od návrhu 1 této metodě dařilo na zkoumaných datových souborech správně odhalit všechna odlehlá pozorování a tím objevit skrytou multikolinearitu.

Závěr

- ▶ Možné problémy s volbou δ.
- ▶ S rostoucím δ se metodě daří správně identifikovat všechna odlehlá pozorování zakrývající multikolinearitu!

6 Plány do budoucna

V tomto posteru byly prezentovány zatím jen úplně základní návrhy, je možné hledat další metody, ať už vycházející z hřebenové regrese či například z regrese s lineárními omezeními, která se také používá při výskytu kolinearity.

První výsledky prezentovaného návrhu 2 se zdají býti velmi slibné, proto je potřeba se touto metodou zabývat podrobněji.

Také zůstává nezodpovězeno mnoho dalších otázek, například v práci nebyl diskutován problém dat, kde je multikolinearita uměle vytvořena odlehlými pozorováními, apod.

e-mail: jurczyk@karlin.mff.cuni.cz; podporováno granty GAČR 201/05/H007 a GAČR 202/06/0408