Posudek oponenta diplomové práce

Název: Metoda upraveného skóre pro lineární model s chybami v regresorech
Autor: Bc. Willy Svoboda

Shrnutí obsahu práce

Diplomová práca študenta, bakalára Willyho Svobodu, sa venuje problematike odhadovania neznámych parametrov v lineárnom regresnom modeli, v ktorom sú niektoré regresory namerané so stochastickou chybou—tzv. *errors-in-variables* modely. Autor v práci predstavuje a diskutuje jeden z možných spôsobov konštrukcie odhadov v takýchto modeloch. Postup je založený na vhodnej modifikácii tzv. skórových rovníc vyplývajúcich z metódy najmenších štvorcov, pričom konkrétna modifikácia závisí na danej špecifikácii uvažovaného regresného modelu a teoretických predpokladov postulovaných v prípade zašumených regresorov.

Z teoretického hľadiska považujem tému za ideálnu pre diplomovú prácu—poskytuje podľa mňa výborný potenciál pre uplatnenie a prípadné ďalšie rozvinutie matematických a štatistických znalosti nadobudnutých počas bakalárskeho a magisterského štúdia na MFF UK. Autor uvádza vo svojej práci niekoľko konkrétnych aj všeobecných modelov a pre uvedené modely sú odvodené príslušné (vhodne upravené) skórové rovnice (i keď z matematického hľadiska sa jedná len o priamočiare parciálne derivovanie a následnú aplikáciu operátoru strednej hodnoty). Teoretické (i.e., asymptotické) vlastnosti získaných odhadov sú stručne zmienené v zmysle neurčitých odkazov na rôzne externé výsledky a zdroje, ale nie je zrejmé, ani ktoré konkrétne výsledky má autor na mysli (napr. akú verziu centrálnej limitnej vety používa), alebo na ktoré zdroje sa odkazuje (absencia akýchkoľvek konkrétnych odkazov).

Z hľadiska obsahu je diplomová práca členená do piatich kapitol. V prvej kapitole autor popisuje klasický lineárny regresný model, metódu najmenších štvorcov a na niekoľkých príkladoch motivuje situáciu s chybne nameranými regresormi. V druhej kapitole je na jednoduchom regresnom modeli predstavený základný princíp fungovania modifikovaných skórových rovníc vychádzajúcich z klasickej metódy najmenších štvorcov. V tretej a štvrtej kapitole je predstavený všeobecný model a tzv. model s validačnou skupinou. Záverečná kapitola sumarizuje výsledky porovnávacej simulačnej štúdie.

Vlastný príspevok autora spočíva v podrobnom až detailnom (matematicky viac-menej korektnom) odvodení upravených skórových rovníc pre jednotlivé uvažované modely a tiež pomerne rozsiahla simulačná štúdia. Z matematického hľadiska ale pôsobi práca trochu rozpačito—chýbajú akékoľvek referencie (`\eqref{}`) na používané matematické výrazy, absentujú odkazy na literatúru, mnohé matematické symboly autor zapisuje foneticky, namiesto symbolicky, občas používa nesprávne alebo nezadefinované značenie, nepresné alebo nejasné odborné formulácie, ale aj preklepy rôznej závažnosti (pár konkrétnych príkladov uvádzam nižšie).

Z formálneho hľadiska je práca na priemernej úrovni. Problém vidím hlavne v nedostatočnej formulácii zmysluplného, logicky korektného a dobre čitateľného, intuitívného textu. Niektoré vety, ale aj celé odstavce sú podľa mňa niekedy nejasné, text sa často a zbytočne opakuje, v niektorých prípadoch si vety dokonca priamo odporujú. Pomerne rozsiahlá simulačná štúdia (viac ako 20 strán), ktorá je úrčite zaujímavým a samostatným autorovým prinosom, by mohla byť prezentovaná výrazne prehľadnejšie a kompaktnejšie. Popis simulácii aj výsledky sú uvedené chaoticky a často neprehľadne (množstvo tabuliek s často nedostatočným popiskom, chýbajúce vysvetlenie zmyslu pri niektorých simuláciach, prípadne chýbajúce ilustrácie, ktoré by určite čitateľovi k celkovému pochopeniu výrazne prospeli). Užitočné by asi bolo aj konzistentne rozlišovať medzi *simulačnou štúdiou*, jednotlivými nezávislými *Monte Carlo opakovaniami*, resp. jednou *konkrétnou vygenerovanou realizáciou*.

Celkovo ale hodnotím prácu ako veľmi zaujímavú, užitočnú z hľadiska teoretického aj praktického využitia, so zrejmým autorovým vlastným prínosom. Zadanie práce považujem za splnené. Vypracovanie a prezentácia je síce na slabšej úrovni, ale napriek všetkému vyššie uvedenému, doporučujem štátnicovej komisii uznať predloženú prácu ako diplomovú prácu na MFF UK.

Stockholm, 24.08.2024

Doc., RNDr., Matúš Maciak, Ph.D.
maciak@karlin.mff.cuni.cz

Otázky k obhajobe

- Na viacerých miestach v práci sa explicitne uvádza, že nejaké konkrétne tvrdenie platí za "*předpokladů uvedené v této práci/kapitole/podkapitole*"... (napr., Věta 1, Věta 2, Věta 3, Věta 6, ale formulácia sa objavuje aj v samotnom texte). Žiadné explicitné predpoklady ale v práci súhrnne uvedené nie sú. Aké konkrétne predpoklady autor vyžaduje pre jednotlivé tvrdenia?

- Na str.11 (a nielen tam) je uvedené, že "*odhad metodou nejmenších čtverců je nestranným odhadem $\boldsymbol{\beta}$, neboť $E[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = \cdots = \boldsymbol{\beta}$.*" Pre akú konkrétnu hodnotu $\boldsymbol{\beta} \in \mathbb{R}^p$ má daný výraz platiť?

- Bude tvrdenie Vety 2 platiť aj pre $\boldsymbol{c} = \boldsymbol{0} \in \mathbb{R}^p$?

- Ktorú konkrétnu Centrálnu limitnú vetu (resp. ktorú verziu CLV) autor využíva pre asymptotickú normalitu výrazu $1/\sqrt{n}\sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\gamma})$ na str.37? Nemalo by asymptotické rozdelenie nejakým spôsobom závisieť aj na hodnote $\boldsymbol{\gamma} \in \mathbb{R}^{2q+5}$?

- Na str.53 autor uvádza, že "*v reálnem světě se můžeme setkat s několika případy, ve kterých může nastat otázka, zda lze asymptotiku ještě využit, či už je natolik špatná, aby se od využití ... raději upustilo*". Akú alternatívnu metódu by autor navrhol využiť v takomto prípade?

- Bolo by možné podobný princíp upraveného skóre uplatniť a využiť aj v prípade odhadovania neznámych parametrov v zobecněných lineárnych regrených modeloch (GLM – Generalized Linear Models)?

Poznámky a pripomienky

- Čo má autor na mysli pod pojmom "*konzistentní vychýlení*" (str.7)? V úvodnej kapitole, ale aj v celej práci je viacej podobných nepresnosti, resp. neúplných formulácii—napr. v regresných modeloch odhadujeme neznáme parametre, nie samotné regresory (str.7); na str.10 nie je zrejmé, ako deterministické vektory $\boldsymbol{x}_i$ a $\boldsymbol{x}_i^j$ súvisia s uvažovaným regresným modelom z Definície 1, resp. s náhodným vektorom $\boldsymbol{X}_i$. Navyše prezentovaná interpretácia modelu navodzuje skôr kauzálnu súvislosť (t.j., podmienené stredné hodnoty sa zmenia, když sa zvýši X_j o jednotku), než pouze asociatívny (t.j., ten správny) vzťah (t.j., rozdiel v podmienených stredných hodnotách prislúchajúci jednotkovému rozdielu v X_j).

- Čo má autor presne na mysli pod pojmom "*zápis po zložkách*"? V práci sa formulácia opakuje niekoľkokrát—napr., zápis

$$\sum_{i=1}^{n} \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) = \boldsymbol{0}_{p,1} \tag{1}$$

  zo str.11, podľa mňa nie je uvedený "*po zložkách*" (ako tvrdí autor). Zápis po zložkách by mal asi vyzerať takto:

$$\sum_{i=1}^{n} X_{ij}(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}) = 0, \quad \text{pre } j = 1, \dots p,$$

  kde $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^\top$; Toto je explicitný zápis j-tej zložky p-rozmerného vektoru v (1). (mimochodom, správne je $\boldsymbol{0}_{p,1}$, resp. $\boldsymbol{X}_i$, namiesto $\boldsymbol{0_{p,1}}$, resp. $\boldsymbol{X_i}$—t.j., index nie je v **bold**)

- Na str.14 autor uvádza, že v modelu "*máme pouze jednu vysvětlující veličinu ... označenou jako Z_i, zatímco druhou vysvětlující proměnnou X_i*" ... V následnej formulácii celkového modelu sa ale objavujú vysvetľujúce premenné tri, X_i, Z_i a W_i; Aký model (resp. koľko regresorov) teda autor uvažuje? Asi by bolo vhodné doplniť podrobnejšiu/presnejšiu formuláciu a aj doplniť značenie, že $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top \in \mathbb{R}^2$ (používa sa $\top$ namiesto T).

- Na Obr.1.1 a Obr.1.2 je (podľa popisku k obrázkom) vykreslený scatterplot pre $Y_i \sim X_i$ a $Y_i \sim W_i$. Samotné grafy ale zakaždým obsahujú os x a os y (ale nie os w).

- Na viacerých miestach by bolo vhodné doplniť odkaz na príslušnú literatúru (napr. odkaz na Berksonův model; odkaz na mnohorozmernú centrálnu limitnú vetu; odkazy v úvode 2.kapitoly, atď.). Celkovo mi príde práca so zdrojmi na pomerne slabej úrovni. V podstate okrem dvoch paragrafov na str.21 autor žiadne zdroje v zbytku práce necituje—na mnohých miestach sa to ale podľa mňa vyslovene žiada.

- Na viacerých miestach by bolo vhodnejšie doplniť formálne presnú (matematickú) formuláciu (namiesto veľmi vágneho a nejasného textu)—napr. v akom zmysle niečo konverguje (v pravdepodobnosti, alebo skoro istě, resp. pre $n \to \infty$), alebo v akom zmysle platia niektoré rovnosti (skoro istě, s pravdepodobnosťou konvergujúcou k jednotke), aké sú rozmery/dimenzie niektorých veličín, prípadne či sa jedná o náhodné, alebo deterministické veličiny, a pod. Taktiež by asi bolo vhodnejšie používať konzistentné značenie pre skóry—niekedy ich autor uvádza ako náhodné kvantity, inokedy ako nenáhodné, niekedy ako funkcie, inokedy ako vektory.

- Čo presne autor testuje v prípade simulačnej štúdie a testovania hypotéz (konkrétne Kapitola 5.5)? Je to nulovosť regresoru (ako to autor explicitne uvádza na str.53), alebo nulovosť príslušného (neznámeho) parametru? V Sekcii 5.5 trochu chýba formálny zápis nulovej a alternatívnej hypotézy. Taktiež nie je zrejmé, pomocou akej testovej štatistiky test funguje a z výsledkov sumarizovaných v tabuľkách 5.10 a 5.11 nie je ani explicitne zrejmé, či je uvažovaná situácia za platnosti nulovej, alebo alternatívnej hypotézy (popisky k tabuľkám by mali byť výrazne podrobnejšie). Aká je empirická sila daného testu? Aké rôzne alternatívy boli uvažované?

- V práci sa podľa mňa výskytuje trochu nadštandardný počet preklepov—jednak vo formulácii nematematického textu (používanie niektorých slov mi dokonca príde vo formálnom odbornom texte ako absolútne nevhodné), ale aj v matematickom značení (jednorozmerné veličiny vs. vektorové kvantity, odhady vs. skutočné (neznáme) parametre, neúplné značenie, a pod.). V tomto smere by si inak celkom zaujímavá a kvalitná práca zaslúžila výrazne dopracovať.