Oponentský posudek diplomové práce

L_1 regrese

Kláry Čelikovské

V práci se studentka zabývá porovnání dvou regresních přístupů. Zatímco, zhruba řečeno, v „klasické“ regresi jde o podmíněnou střední hodnotu a od toho se odvíjí i použití kvadratické ztrátové funkce, v pro L_1 regresi je typické použití absolutní hodnoty a odhadujeme podmíněný medián, případně jiné kvantily.

Autorka se pro standardní model $Y = \boldsymbol{X}^T\beta+\varepsilon$ zabývá dvěma hlavními úkoly. První je asymptotické chování odhadu a druhý je porovnání L_1 a L_2 regrese. Za vlastní přínos autorky lze považovat rozepsání některých kroků důkazu asymptotické normality odhadu regresních koeficientů, které jsou v původním článku Basseta a Koenkera z roku 1978 jen naznačené, nebo prohlášené za zřejmé. Autorka také bere vysvětlující proměnné důsledně jako náhodné vektory, což je chvályhodné, neboť ve zdrojovém článku není explicitně řešeno, zda $\boldsymbol{X}$ je náhodný vektor, nebo jde o pevné konstanty. Dalším vlastním přínosem je simulační studie ukazující některé slabé stránky L_2 regrese v případě, že náhodné chyby nejsou normálně rozdělené, případně jsou nesymetrické či s těžkými chvosty.

Na můj vkus je práce stručná. Nepochybuji o tom, že doplnění důkazu věty 7 dalo dost práce. Simulační studie je také stručná a zde se přitom nabízela možnost studovat i takzvané body zlomu, tedy situace, kdy změna jediného či několika málo pozorování zcela změní odhad regresních koeficientů. Přitom autorka mohla při simulacích využít bohatou nabídku knohoven v R, rozšíření simulací proto nemuselo ani být časově náročné. Zde bych se rád dozvěděl, proč jsou simulace omezeny jen na relativně jednoduché případy a jejich ilustrační potenciál tak nebyl využit. Jinak práci vidím jako přínosnou a zajímavou a při jejím psaní se jistě autorka naučila i znalosti, které přesáhly rámec předmětů vyučovaných na oboru PMSE.

Několik dotazů a připomínek k práci:

(1) Je předpokládáno, že náhodná veličina ε a náhodný vektor $\boldsymbol{X}$ jsou nezávislé?

(2) Strana 5: Proč není třeba v lemmatu 1 předpokládat konečnost střední hodnoty X?

(3) Strana 6 nahoře: Ve vzorci pro $M(\theta)$ jsou některá znaménka špatně. Navíc je zde tiše předpokládáno, že distribuční funkce F je spojitá, což ale v předpokladech lemmatu není. Znamená $M'(\theta_-)$ derivaci zleva?

(4) Strana 8: Není zde dvakrát předpokládáno totéž? Nejprve je formulován předpoklad $Y = \boldsymbol{X}^T\beta + \varepsilon$ s tím, že $F_\varepsilon^{-1}(\alpha) = 0$. Po definici odhadu je řečeno „pokud platí $F^{-1}_{Y|\boldsymbol{X}}(\alpha) = \boldsymbol{X}^T\beta$“, ale to podle mého plyne již z předchozího předpokladu.

(5) Strana 10 zcela nahoře: je v maximalizaci správně $+$, nebo $-$?

(6) Strana 11 nahoře: Při definici funkce $\boldsymbol{\zeta}$ je matoucí ji psát jako vektor $n-k$ funkcí $(\zeta_1, \ldots, \zeta_{|\overline{H}|})$. Zároveň je špatně uvedený obor hodnot této funkce, má být $\mathbb{R}^k$.

(7) Lemma 4: Předpoklad $\boldsymbol{\zeta}(H, \boldsymbol{v}) \in C[\boldsymbol{0}, 1]$ má platit skoro jistě. Co to přesně znamená pro $\boldsymbol{X}$, Y a H?

(8) Důkaz lemmatu 4: Hned na začátku důkazu je neostrá nerovnost ≥ 0, ale mluví se o kladných přírůstcích ve všech směrech. Přitom tato nerovnost je triviálně splněna funkcí konstantní na okolí $\boldsymbol{a}$. Proč je v absolutní hodnotě zachováno znaménko $-$?

(9) Věta 7: předpoklad (iii) je $\|\mathbb{X}_n\|_\infty = o(\sqrt{n})$. Matice $\mathbb{X}$ je náhodná, v jakém smyslu tedy toto malé o je?

(10) Důkaz věty 7: Značení u_i se vyskytuje ve dvou významech, jednak jako reziduum $Y_i - \boldsymbol{X}_i\widehat{\beta}$ (zde navíc chybí transpozice u $\boldsymbol{X}_i$) a pak jako složka vektoru $\boldsymbol{u}(H) = \boldsymbol{Y}(H) - \mathbb{X}(H)\beta$.

(11) Strana 17: Je někde ukázáno (2.12)?

(12) Strana 20: V odvození rozptylu $\boldsymbol{z}_i(\boldsymbol{\delta}, H)$ se odečítá vektor od matice. Odečítaný člen by měl být doplněný o transpozici první části a výraz v závorce umocněný na druhou.

(13) Strana 21: nemá být na řádku 3 místo mocniny $1/2$ spíš $-1/2$?

(14) Věta 8: Co přesně znamená v bodě (ii), že existuje pevné γ takové, že $\beta_2 = \gamma/\sqrt{n}$ pro všechny velkosti výběru n? V důkazu potom vidíme $\gamma = \sqrt{n}\beta_2$ a toto γ je střední hodnotou limitního normálního rozdělení. Znamená to, že $\beta_2 \to 0$?

(15) Strana 24: Proč je výhodné nahradit odhad $\mathbb{Q}_{22}$, když musíme stejně odhadovat $\mathbb{Q}_{12}$, $\mathbb{Q}_{21}$ a $\mathbb{Q}_{11}$?

(16) Strana 24: na devátém řádku odspodu je „ekvivalentní“ odhad kvantilové funkce, ale vypadá trochu podezřele, pokud $\mathbb{I}$ je indikátor.

(17) Strana 29: při popisu simulací je udedeno, že je vygenerován vektor pozic odlehlých pozorování (bod 1). Odlehlých pozorování čeho?

(18) Strana 30: V modelu F_2 jsou až neuvěřitelně velké hodnoty MSE pro L_2 regresi. Je-li $\beta_1 = 1$, pak by odhady β musely být v rozmezí $(-1000, 1000)$. Je tomu tak, opravdu pro 100 pozorování vycházely tak velké odhady? Dále je zde zajímavé, že MSE pro 100 pozorování je větší než pro 50 i než pro 500 a to řádově v případě β_1. Máte pro to nějaké vysvětlení? Podobná zvláštnost se vyskytuje i v tabulce 3.5, jen opačně. MSE je nejmenší pro $n = 100$ a prudce vzroste pro $n = 500$.

Práce jistě splnila zadání a přes uvedené často formální výtky ji považuji za vyhovující standardům diplomové práce. U obhajoby doporučuji zdůraznit vlastní vklad autorky a vysvětlit ty z výše uvedených bodů, které nelze brát za překlep či formulační nedostatek.

Doporučuji tuto práci **uznat za diplomovou práci** pro obor Pravděpodobnost, matematická statistika a ekonometrie.

Daniel Hlubinka
Ve Zbraslavi 26.6.2020