Posudek vedoucího na bakalářskou práci

Autorka bakalářské práce: Ksenia Molodkina

Název bakalářské práce: Krokové metody v lineární regresi a jejich vlastnosti

Vedoucí bakalářské práce: Mgr. Milan Bašta, Ph.D.

V teoretické části bakalářské práce je zaveden pojem (normální) lineární regresní model. Jsou studovány vlastnosti odhadu parametrů tohoto modelu metodou nejmenších čtverců. V dalších sekcích teoretické části jsou zavedeny metody sestupného a vzestupného výběru nezávisle proměnných a metoda krokové regrese.

V praktické části bakalářské práce jsou prostřednictvím Monte Carlo simulací studovány vlastnosti sestupného a vzestupného výběru nezávisle proměnných a metoda krokové regrese, a to konkrétně vychýlení odhadů regresních parametrů ve výsledném modelu, četnost detekce správné konfigurace regresních parametrů a predikční schopnost výsledného odhadnutého modelu. Autorka prezentuje výsledky simulací v grafech a tabulkách a předkládá i kvalitativní vysvětlení získaných výsledků.

V kontextu teoretické části bakalářské práce je třeba ocenit, že studentka pracovala s maticovým zápisem modelu, porozuměla tomuto zápisu a prostřednictvím tohoto zápisu dodala i několik důkazů vlastností odhadu regresních parametrů metodou nejmenších čtverců. Teoretická část práce obsahuje jen málo skutečně závažných formálních chyb a nedostatků. V kontextu praktické části bakalářské práce zase oceňuji, že si autorka v softwaru R samostatně naprogramovala vlastní funkce, které sestupný, vzestupný a krokový výběr nezávisle proměnných provádějí. Obdobně, si vytvořila i další funkce, které vyhodnocují požadované vlastnosti výše uvedených metod výběru nezávisle proměnných. Zdrojové kódy všech těchto funkcí jsou přiloženy jako součást elektronické přílohy bakalářské práce. Uveden je též i podrobný komentář příslušného kódu. Výsledky Monte Carlo simulací se zdají vesměs rozumné a jejich kvalitativní vysvětlení se jeví většinou (ale ne vždy, viz níže) jako korektní.

Jako nedostatky práce je možné uvést následující: Teoretická i praktická část práce jsou relativně nepřehledné a čtenář se místy hůře orientuje v tom, kam autorka směřuje. Obsah některých pasáží je místy „duplikován“ (viz např. některé pasáže v kapitole 1.1). Některé formulace jsou nesrozumitelné (viz např. text následující ihned pod rovnicí (36)), nesprávné (např. o vektoru vyrovnaných hodnot by bylo lepší hovořit jako o odhadu střední hodnoty vektoru $\boldsymbol{Y}$, a nikoliv jako o odhadu $\boldsymbol{Y}$, viz např. Definice 9) nebo nepřesné/vágní (viz např. str. 22, dole: "... generuje regresní modely, počítá a vrací charakteristiky vytvořených regresních modelů ...", zde by bylo korektnější psát "... charakteristiky odhadnutých regresních modelů ... "; nebo tvrzení ze str. 23, že S_{sim} je výběrová směrodatná odchylka pro soubor simulovaných dat). Popis toho, jak přesně byly modely v rámci Monte Carlo simulací generovány, je místy těžkopádný (např. ze str. 23 je jen stěží pochopitelné, že funkce uvedené na této straně slouží k tomu, aby se měnil vždy jeden parametr z parametrů Tabulky 2, zatímco ostatní parametry této tabulky zůstávaly neměnné). Značení není místy korektní (např. ve Větě 5 není vektor $\boldsymbol{\beta}$ uveden tučně v souladu s předchozím značením) a použitá terminologie není občas optimální (např. na str. 24 je mluveno o vychýlení parametrů ve vektoru $\widehat{\boldsymbol{\beta}}$). Odkazy na některé rovnice nejsou uvedeny ideálně. Např. na str. 9 je před rovnicí (10) odkazováno na rovnici (7), lepší by však bylo odkázat se na rovnici (9). Obdobně, viz odkaz na rovnici (6) na začátku důkazu Věty 1 nebo odkaz na kapitolu (2.6) na str. 15, jež vůbec neexistuje.

Jako příklad nekorektního vyhodnocení závěru Monte Carlo simulací je možné uvést tvrzení na str. 29, kde je uvedeno, že "hodnoty relativních četností klesají s rostoucími hodnotami parametrů PtoStay a PtoEnter", což je však v rozporu s obr. 5. Dalšího případu „podezřelých“ výsledků se týká má 1. otázka níže. Dále, se v textu vyskytuje snaha o zobecnění výsledků Monte Carlo simulací na situace, které nebyly v rámci těchto simulací studovány (viz např. sekce 2.3.4, kde jsou výsledky studia vychýlení odhadů prezentovány ve vší obecnosti, ačkoliv ve vlastní analýze byla použita jen jedna konkrétní konfigurace vektoru regresních parametrů). Dalším příkladem „přílišného“ zobecňování výsledků jsou některá tvrzení ze str. 31, str. 32, str. 33. nebo str. 37.

V textu místy chybí slova, občas je používán nekorektní slovosled nebo se vyskytují překlepy. Autorka se nevyvarovala ani hrubých gramatických chyb.

Celkově lze shrnout, že Ksenia Molodkina prokázala ve své bakalářské práci schopnost pracovat s obtížnějším matematickým aparátem, naprogramovat vlastní skripty pro provádění Monte Carlo simulací a také schopnost vyhodnotit výsledky těchto simulací a kvalitativně je vysvětlit. Autorce se však ne vždy podařilo napsat text v úplně srozumitelné a čtenářsky příjemné formě. Je také vidět, že některé pasáže jsou poněkud nedotažené, s nekorektními tvrzeními a text je místy neformální až nepřesný. Rezervy je též možné shledat v práci s literaturou, neboť v práci nejsou uvedeny fakticky žádné výsledky předchozích studií týkající se vlastností krokových metod (ačkoliv takové studie existují a byly autorce poskytnuty). Závěrem lze říci, že autorka cíle

bakalářské práce splnila, některé části práce však mohly být vypracovány pečlivěji, dotaženy více do konce a práce jako celek mohla být více vyladěna.

Na autorku bych měl následující dotazy (řadím je podle důležitosti):

1.) Vysvětlete, jak je možné, že ve sloupci "Plný model" v Tabulkách 3, 4 a 5 jsou pro stejné poměry σ^2/n vždy stejné údaje (ačkoliv vektor regresních parametrů je různý). Nejedná se o chybu v přepisu či vyhodnocení výsledků? Pokud to chyba v přepisu a vyhodnocení výsledků není, vysvětlete proč ne. Pokud se jedná o chybu, okomentujte, zda zůstávají Vaše tvrzení ze sekce 2.3.3 v platnosti i pokud byste tuto chybu opravila.
2.) Na str. 23 mluvíte o tom, jak vyhodnocujete přesnost výsledků Monte Carlo simulací. Vysvětlete, odkud vzorec (SE), který pro tyto potřeby využíváte, plyne. Detailně vysvětlete aplikaci tohoto vzorce pro získání čárkovaných čar na Obrázku 1. (Ve vlastním textu je totiž vysvětlení vzorce „SE“ velmi neformální a vágní.)
3.) V Závěru (str. 38) píšete, že "Z tohoto důvodu z hlediska testů hypotéz je lepší preferovat plný model". Vysvětlete přesněji, co tím máte na mysli.
4.) V teoretické části textu často mluvíte o pozitivně definitní či pozitivně semidefinitní matici. Vysvětlete, co tyto pojmy znamenají.

24. srpna 2014,

Mgr. Milan Bašta, Ph.D.
Katedra statistiky a pravděpodobnosti, FIS, VŠE, Praha