Posudek oponenta bakalářské práce

Název: Krokové metody v lineární regresi a jejich vlastnosti
Autor: Ksenia Molodkina

Shrnutí obsahu práce

Práce pojednává o krokových metodách výběru proměnných v lineárním modelu. Jak již avizuje zadání práce, jedná se o metody používané spíše v minulosti (s ohledem na současnou dostupnost metod mnohem lepších). Jedním z cílů práce proto bylo prozkoumat vlastnosti krokových metod (a následně dojít k závěru, že jejich použití obvykle není příliš vhodné). Práce je rozdělena do dvou hlavních kapitol nazvaných trochu nešťastně jako „Teoretická část“ a „Praktická část“. Zavedenou praxí u bakalářských prací psaných na MFF UK na programu Matematika je přitom nazývat kapitoly spíše dle jejich obsahu a nesnažit se striktně oddělovat „teorii“ a „praxi“. Tak zvaná teoretická část se zejména věnuje shrnutí základních vlastností lineárního modelu a vysvětlení algoritmů krokových metod výběru proměnných. Tak zvaná praktická část se zabývá (a) popisem programových procedur vytvořených v rámci bakalářské práce, (b) popisem a shrnutím výsledků simulační studie.

Celkové hodnocení práce

Téma práce. Téma se mi jeví jako přiměřené pro bakalářskou práci. Zpracování víceméně splňuje zadání práce. S ohledem na požadavek „budou analyzovány a diskutovány nežádoucí vlastnosti těchto [krokových] metod“ bych však uvítal též alespoň drobné teoretické studium krokových metod, resp. důsledků nezahrnutí důležitých regresorů, resp. zahrnutí nadbytečných regresorů, která při použití krokových metod často nastávají. V rámci předložené práce je vše studovány prakticky pouze pomocí simulací.

Vlastní příspěvek. Za vlastní příspěvek autorky lze nepochybně považovat druhou kapitolu práce („Praktická část“), v rámci které autorka samostatně implementovala krokové metody v programu R (jednotlivé procedury jsou stručně popsány též v bakalářské práci) a provedla a shrnula simulační studii s cílem prozkoumat vlastnosti krokových procedur.

Matematická úroveň. Matematická úroveň práce je spíše podprůměrná. V rámci první kapitoly („Teoretická část“) se autorka snaží formou definic a vět shrnout základní vlastnosti lineárního modelu, jehož parametry jsou odhadovány metodou nejmenších čtverců. Mnohá vyjádření jsou však značně nepřesná, někdy i chybná (viz připomínky níže).

Práce se zdroji. Autorka cituje použité zdroje. Práce dle mého názoru neobsahuje doslova zkopírované ani otrocky přeložené pasáže. Vadou na kráse je však nekonzistentní způsob odkazů na literaturu. Na některých místech je používán styl *(Autor, Rok)* (např. str. 12), na jiných místech potom styl *[číslo odkazu]* (např. str. 15).

Formální úprava. Po formální stránce se práce na první pohled jeví v pořádku. Hrubé gramatické chyby se v práci vyskytují v míře malé. Některá stylisticky ne úplně nejšťastnější vyjádření lze prominout s ohledem na fakt, že čeština (ve které je práce psána) patrně není mateřským jazykem autorky. Naprosto zbytečnou vadou jsou však anglicky psané nadpisy sloupců v tabulkách (viz např. str. 15–18) i v legendách obrázků (viz např. str. 24).

Obecnější připomínky

1. Data použitá v oddíle 1.6 jsou naprosto nevhodná ve spojitosti s klasickým lineárním modelem. Dokonce bych řekl, že použití lineárního modelu v tomto kontextu je dosti hrubou chybou. Domnívá se snad autorka, že se lineární model hodí pro **binární** odezvu, kterou je uvažovaná remise rakoviny?

2. V první kapitole se operuje s maticí modelu, která je nenáhodná. Toto usuzuji např. z vyjádření „Vektor $\mathbb{X}\boldsymbol{\beta}$ je nenáhodný.“ na str. 8 a dále z průběhu některých důkazů. Simulační studie je však prováděna s náhodnou maticí modelu (jednotlivé řádky jsou generovány z vícerozměrného normálního rozdělení). Proč tento (zbytečný) rozpor mezi první a druhou kapitolou?

3. Autorka vynaložila poměrně velkou snahu popsat obšírně výsledky simulační studie. Výsledkem je však, dle mého názoru, poměrně nepřehledný text, ve kterém zapadají ta nejdůležitější sdělení. Taktéž mi vadí, že je na začátku druhé kapitoly popsána jistá sada simulačních scénářů, nicméně v dalších jednotlivých částech druhé kapitoly jsou potom popisovány výsledky vždy jenom pro různě vybírané scénáře. Uvítal bych shrnutí všech výsledků ve formě (vhodně formátovaných) tabulek či obrázků a potom slovní popis toho nejdůležitějšího.

Konkrétnější připomínky

U některých z níže uvedených připomínek se možná jedná o pouhé překlepy. Ne vždy jsem však schopen posoudit, zda se jedná opravdu jenom o překlep a kdy o nepochopení studovaného problému. Níže uvedený výčet není zcela jistě úplný.

1. Zavedení lineárního modelu (Definice 1 a následný text na str. 8) je dosti chaotické. Kupříkladu reziduální rozptyl je zaveden jenom tak mezi řečí v půlce str. 8.

2. Vztah (7) na str. 9 není roven vztahu (8), jak se autorka snaží tvrdit. Je třeba rozlišovat mezi řešenou soustavou a jejím řešením!

3. V začátku důkazu Věty 1 na str. 9 by spíše mělo být, že daný vztah plyne z (8) a nikoliv z (6).

4. Na začátku str. 10 autorka píše, že $\widehat{\boldsymbol{Y}}$ může být považováno na nejlepší aproximaci vektoru $\boldsymbol{Y}$ vzniklou pomocí lineárních kombinací sloupců matice $\mathbb{X}$. Nikde však není řečeno, v jakém smyslu „nejlepší“.

5. V definici 9 na str. 11 (a potom na některých místech dále) se operuje s výrazem „lineární odhad“. Toto není nikde definováno a v daném místě je též poměrně obtížné si domyslet, co oním lineárním odhadem autorka myslí. V téže definici autorka též opomněla sdělit, co je $\widehat{\boldsymbol{Y}}$. Pokud před tím definované vyrovnané hodnoty, tj. $\widehat{\boldsymbol{Y}} = \mathbb{H}\boldsymbol{Y}$, potom je celá definice dosti nesmyslná, neboť s $\boldsymbol{Y}$ se zde operuje, jako kdyby se jednalo o odhadovaný (nenáhodný) parametr a nikoliv náhodný vektor reprezentující odezvu, jehož (lineární) funkcí je $\widehat{\boldsymbol{Y}}$.

6. Není pravda, že $\boldsymbol{\varepsilon} = \mathbf{0}$ v důkaze Věty 3 na str. 11.

7. Není pravda, že $\widehat{\boldsymbol{Y}}$ je nejlepším možným lineárním odhadem vektoru $\boldsymbol{Y}$, jak se tvrdí na konci důkazu Věty 3 na str. 12. Tato již druhá záměna něčeho, co má býti odhadováno s náhodným vektorem (daty), pomocí něhož má ono něco býti odhadnuto, ve mně vyvolává dojem, že autorka nemá v celé věci úplně jasno a libovolně zaměňuje náhodný vektor odezvy $\boldsymbol{Y}$ s jeho střední hodnotou.

8. V popisech konečných modelů (vztahy 33–35 na str. 16–19) se mi dosti nelíbí rovnítko kladené mezi odezvu a lineární prediktor (odhad **střední hodnoty** odezvy).

9. Domnívám se, že ME (výraz 38 na str. 23) spíše vyjadřuje (střední čtvercovou) chybu odhadu střední hodnoty odezvy a nikoliv chybu predikce (která by musela vzít do úvahy též variabilitu náhodné složky lineárního modelu). Taktéž mi vadí, že není řečeno, co se myslí **chybou**. Až z kontextu usuzuji, že se jedná o střední čtvercovou chybu.

Otázky

1. Prosím o přesnou definici ME zavedeného na str. 23 a následné odvození vztahu (38).

2. V rámci simulačních studií počítáte nejrůznější intervaly spolehlivosti (např. pro vychýlení odhadů, viz Obrázek 1). Prosím o vysvětlení, jakým způsobem jsou tyto intervaly spolehlivosti získány?

3. Kromě několika dalších nechápu větu: „Chyby, které vznikají při testování hypotéz, ovlivňují taky výslednou konfiguraci odhadnutých paranetrů. Z tohoto důvodu z hlediska testů hypotéz je lepší preferovat plný model.“ uvedenou ve třetím odstavci Závěru na str. 38. Prosím o vysvětlení významu tohoto sdělení.

Závěr

Práci považuji za spíše podprůměrnou (zejména s ohledem na poměrně vysoké množství chyb a hrubších nepřesností v matematických pasážích) nicméně **vyhovující** a **doporučuji** ji uznat jako bakalářskou práci.

doc. RNDr. Arnošt Komárek, Ph.D.

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze

V Praze 19. srpna 2014