

Oponentní posudek diplomové práce

Jméno a příjmení autora práce: Bc. Bohuš Nemčovič

Název diplomové práce: Aplikácia metódy bootstrap pri posúdení kvality lineárneho regresného modelu

Zvolené téma

Zvolené téma považuji za zajímavé (až atraktivní) a pro diplomovou práci „statistického“ oboru za velmi vhodné. Název práce podle mého názoru příliš dobře neodráží, co je obsahem práce samotné (porovnání různých přístupů k odhadu vybraných charakteristik vyjadřujících kvalitu regresního modelu prostřednictvím simulační studie). Přestože sám autor svou práci jednoznačně „neklasifikuje“ (ve smyslu pokynů pro zpracování diplomové práce), jedná se o práci usilující o teoretický přínos k současnému stavu poznání.

Množství a kvalita odvedené práce

Je evidentní, že autor své diplomové práci věnoval značné úsilí. Provedená analýza jde nad rámec znalostí, které si student mohl osvojit v povinných předmětech v rámci svého magisterského studia. Bohužel zde před dalším musím konstatovat, že si nejsem zcela jist vlastní dostatečnou kompetencí k náležitému posouzení předložené práce a je tak možné, že některé mé komentáře či otázky níže mohly vyplynout z nepochopení, které nelze přičítat textu diplomové práce samotné.

V práci je mj. k dispozici čtenáři několik obrázků a tabulek prezentujících výsledky dosažené v provedených simulacích. Celkový rozsah diplomové práce je standardní. V práci se vyskytuje jen úplné minimum překlepů.

Použitá literatura a čerpání z ní

Pozitivní vlastností práce je struktura použitých zdrojů literatury, ve kterých dominují zdroje v angličtině.

Úroveň zpracování

Po stránce obsahové má práce logickou strukturu odpovídající typu předložené práce, kdy je vcelku jednoznačně oddělena teoretická (kapitoly 1 a 2) a simulační část (kapitola 3).

Je škoda, že práce neobsahuje seznam použitých symbolů. Autor práce zřejmě přejímal pro různé části své práce symboliku z různých zdrojů literatury, což je sice pochopitelné, ale orientaci v textu práce to nezjednodušuje: některé z použitých symbolů nejsou vysvětlené (popsané) vůbec, některé symboly v práci postupně označují něco jiného. Poněkud nevyužitý potenciál vidím také v relativně skromném popisu (komentářích) dosažených výsledků (obrázky a tabulky).

Splnění cílů práce

V úvodu práce jsou definovány následující cíle:

- ukázat, že po aplikovaní procedúr na výstavbu lineárneho regresného modelu už klasické odhady smerodatnej odchylky určitých odhadov nie sú presné
- posúdiť do akej miery jednotlivé procedúry znižujú presnosť odhadu smerodatnej odchylky odhadu regresnej funkcie v bode alebo odhadu regresného parametru a ukázať pomocou metódy bootstrap možnosť opravy

Výše uvedené (obecně formulované) cíle (záměry) hodnotím jako značně ambiciózní, možná až příliš (viz také níže). Oba cíle nicméně považuji nejméně za částečně splněné

Vlastní přínos autora

I přes některé výše (a níže) uvedené výtky hodnotím práci jako celek pozitivně. V teoretické části práce propojuje symboly a pojmy z regresní analýzy a z metody bootstrap, ve 3. kapitole je pak čtenář seznámen s množstvím výsledků (prezentovaných v grafické i v tabulkové podobě), kterých autor dosáhl v provedených simulacích (v programu R). Diplomová práce je zajímavá a o značném přínosu samotného autora není žádných pochyb.

Odborná úroveň

Zřejmě mj. v důsledku kombinace různých zdrojů literatury a celkové obtížnosti zvoleného tématu jsou některé části práce více či méně nesrozumitelné. Zejména jsem velmi tápal, co autor chápe pod pojmem „model“ a co má na mysli pod označeními „skutečný model“, „správný model“, „odhadnutý model“, „finální model“ atd. Čtenář si relativně často klade otázky „Proč?“, „Čeho?“ (skutečný model, směrodatné odchylky, rozptyl), „V jakém smyslu?“ (je neuspokojivější, je nejvhodnější), „Jakých?“ apod.

V dobrém slova smyslu zdůrazňuji práci v programu R (části kódů jsou k dispozici v textu práce).

Interpretace výsledků, srozumitelnost, výstižnost formulací

Předložená práce je bohužel poněkud úsporná, pokud jde o interpretaci výsledků ve stěžejní Kapitole 3. Text celé práce byl pro mě v některých pasážích relativně obtížně čitelný (opakuji, že se může jednat o mou nedostatečnou kompetenci pro porozumění textu a nikoliv o nedostatek předložené práce), stylisticky dobře strukturovaný.

Formální úprava práce

Po formální stránce (úprava) předložené práci moc nemám, co vytknout. Práce obsahuje jen minimum překlepů.

Závěr

Práce se zabývá velmi obtížným tématem a klade si ambiciózní cíle. Autor se v práci věnoval porovnání několika způsobů (Monte Carlo simulace, „naivní“ odhad, bootstrapový odhad) odhadu odmocniny ze střední čtvercové chyby odhadu hodnoty regresní funkce ve vybraném bodě a odhadu odmocniny ze střední čtvercové chyby odhadu vybraného regresního parametru. Práce nejméně částečně splnila vytčené cíle a pro čtenáře představuje zajímavý zdroj informací. Cíle nebyly (a domnívám se, že ani nemohly být) dosaženy zcela, protože autor ve svých simulacích vychází z „fixního“ počtu proměnných, z konkrétních rozdělení vysvětlujících proměnných, z konkrétního počtu pozorování atd. Za pozitivum lze jednoznačně označit zajímavé výsledky provedených simulací prezentované v Kapitole 3 a také některé aspekty teoretického výkladu vybraných metod. Pro účely obhajoby práce před komisí navrhuji (vybírám z velkého množství možných) následující otázky/podněty:

- 1) Na str. 4 (v kapitole 1.1) je uvedeno (mj.), že předpokládáme, X je nenáhodná vysvětlující proměnná. Totéž je zřejmě předpokládáno v kapitole 1.2. Ve všech simulacích jsou nicméně vysvětlující proměnné uvažované jako náhodné veličiny. Vysvětlíte tento (zdánlivý?) rozpor.
- 2) Co ve Vaší práci značí symbol y_i ? Hodnotu vysvětlované proměnné pro i -té pozorování (str. 4) nebo „náhodnou proměnnou“ (str. 8)?
- 3) Co máte na mysli „správnou specifikací modelu“ na str. 9? Dle textu pod vzorcem 1.39 se zřejmě jedná o něco jiného, než splnění slabé sady předpokladů.
- 4) Bohužel příliš nerozumím např. výroku na str. 11 „V běžné praxi bohužel často není známý skutečný model a je nutné tento vztah vyhledat pomocí dat, která máme k dispozici.“ Co zde máte na mysli „skutečným modelem“?
- 5) Provedené simulace a dosažené výsledky jsou založené na tom, že (odhad?) RMSE získaný pomocí metody Monte Carlo představuje „skutečnou“ RMSE. V komentářích k výsledkům

prezentovaným v tabulce 3.12 (str. 43) je nicméně konstatováno, že „je jasné vidět selhání Monte Carlo odhadu“. Zkuste prosím vysvětlit/obhájit, jak je tedy možné uvažovat o hodnotě RMSE získané metodou Monte Carlo jako o „skutečné“ hodnotě RMSE, když i tato metoda může „selhat“?

- 6) Zhruba polovina provedených simulací se věnuje odhadu regresního parametru β_1 . Vzhledem k tomu, že v části práce zkoumající odhad RMSE odhadu tohoto parametru je použita (zkoumána) i transformace DAF3 (automatizovaný výběr vysvětlujících proměnných; minimalizace AIC), nejsem si bohužel jist, o jaký parametr (parametr čeho) se vlastně jedná (zřejmě se nejedná o dílčí parametr vysvětlující proměnné X_1 , protože ta zřejmě při DAF3 vůbec nemusela v modelu být)?
- 7) Na str. 23 uvádíte, že $\beta_0=0$ a $\beta_i=1$, z čehož by se mj. mohlo zdát, že v práci jsou uvažovány pouze modely bez absolutního členu. Toto je ale v (zdánlivém?) rozporu s ostatním textem práce. Prosím vysvětlíte.
- 8) Ve svých simulacích (pochopitelně) vycházíte z mnoha předpokladů ($n=50$, $p=5$, vysvětlující proměnné jsou náhodné veličiny s normovaným normálním nebo rovnoměrným rozdělením). K čemu jsou s ohledem na tuto skutečnost Vaše výsledky z praktického hlediska dobré/užitečné?
- 9) Proč je ve vzorci 3.7 dvojnásobek a nikoliv druhá mocnina?
- 10) Co vyjadřují 3 hodnoty v řádku „100“ pro jednotlivé počty dat v sloupcích „Monte Carlo“ a „Naivní“ v tabulce 3.2 (str. 27)?
- 11) Jak/čím byste vysvětlil dvojvrcholové rozdělení, které se ve Vašich výsledcích několikrát objevilo u „Naivního“ odhadu?
- 12) Jak je možné, že „Bez DAF“ a v případě „DAF3“ je selhání (rozdíl v hodnotách mediánů RMSE) větší při větším počtu pozorování (viz tabulka 3.13 na str. 44)?

Vzhledem ke všemu výše uvedenému práci hodnotím jako poněkud rozporuplnou (nesmírně ambiciózní cíle a obtížné metody, ale také spousta nejasností v předloženém textu). S přihlédnutím ke všem výše zmíněným skutečnostem práci nicméně jednoznačně doporučuji k obhajobě a s ohledem na průběh obhajoby navrhuji celkové hodnocení *výborně až velmi dobře*.

V Praze dne 18.8.2018
Ing. Karel Helman, Ph. D.
KSTP VŠE Praha