

Posudek vedoucího na diplomovou práci

Autor diplomové práce: Bc. Bohuš Nemčovič

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

Cílem diplomové práce je pomocí Monte Carlo simulací posoudit vliv data analytických funkcí (detekce odlehklých a vlivných pozorování, transformace vysvětlované proměnné, výběr vysvětlujících proměnných do modelu) použitých při výstavbě lineárního regresního modelu na validitu vztahů plynoucích z klasické teorie lineární regrese pro odhad odmocniny ze střední čtvercové chyby (RMSE) odhadu regresního parametru a pro odhad RMSE odhadu regresní funkce. Paralelním a neméně důležitým cílem je použití metody bootstrap pro odhad příslušných RMSE a posouzení vhodnosti této metody.

Vzhledem k tomu, že vybrané téma a vytyčené cíle vyžadují dobrou znalost klasické teorie lineární regrese, metody bootstrap, metody Monte Carlo a také schopnost příslušné postupy naprogramovat, lze dané téma považovat za nadprůměrně náročné. Z teoretické části (tj. z kapitoly 1 a 2) je zřejmé, že student dané problematice rozumí. Většina textu v této části – co se týče obsahu – a také většina uvedených vztahů a značení je správných. Výsledky v praktické části (tj. v kapitole 3) se jeví také jako rozumné a možné (až na pár drobných nejasností). Oceňuji také naprogramování vlastního kódu simulace v softwaru R. Celkově hodnotím diplomovou práci pozitivně. Níže uvádím nedostatky, kterých jsem si v práci všimnul. Tyto nedostatky (byť by odstranění některých z nich diplomové práci jednoznačně prospělo) však nesnižují celkový pozitivní dojem z diplomové práce.

Text teoretické části je psaný na některých místech poněkud těžkopádněji a struktura výkladu je místy trochu nepřehledná. Např. v sekci 1.2.2 autor představuje vlastnosti odhadů metodou nejmenších čtverců: tyto vlastnosti jsou prezentovány v dané sekci jak pro model jednoduché lineární regrese, tak i pro model vícenásobné lineární regrese, a to takovým způsobem, který je místy neuspořádaný. Ve stejné sekci je obdobně poněkud nepřehledně, a aniž by bylo odděleno od předchozího obsahově poněkud jiného textu, vložena také pasáž o střední čtvercové chybě odhadu (MSE) nebo povídání o data analytických funkcích. Dle mého názoru by bylo vhodnější tyto různé aspekty výkladu strukturovat do oddělených sekcí či podsekcí. Obdobně trochu těžkopádnější výklad je např. i v sekci 1.3.1. Přehlednější strukturování výkladu do samostatných odstavců, sekcí či podsekcí a obratnější vyjadřování by bylo vhodné i v některých pasážích praktické části diplomové práce, například na začátku Kapitoly 3 (Simulace).

Poslední odstavec sekce 1.2.2 je pro čtenáře v dané fázi čtení diplomové práce relativně těžko srozumitelný. Nesrozumitelná je zřejmě také definice naivního odhadu uvedená pod rovnicí 1.53. Není totiž zřejmé, že pisatel má pod naivními odhady na mysli odhady *RMSE* odhadu, resp. odhady *směrodatných chyb odhadu*, a nikoliv odhady parametrů. Definice outlierů v půlce strany 11 také není úplně korektní a není plně v souladu s následující korektnější definicí ze sekce 1.3.1. V sekci 1.3.1 není také zřejmé, proč autor za účelem daného výkladu uvažuje konkrétní formu modelu uvedenou v rovnici 1.63. Není též úplně jasné, proč jsou vlivná pozorování v kontextu DAF1 vypouštěna (vlivné pozorování totiž nemusí být automaticky chybné pozorování, resp. outlier). Odhad RMSE (odhadu regresního parametru, resp. odhadu regresní funkce) s využitím bootstrapu z konce sekce 2.2 není příliš srozumitelný. Obdobně i v praktické části jsou některé pasáže, které by vyžadovaly lepší vysvětlení: Např. proč jsou hodnoty vysvětlujících proměnných ve vanila modelu generovány z normálního rozdělení, zatímco v jiných modelech jsou generovány z rozdělení rovnoměrného? Detailnější komentář by také asi zasluhoval popis Tabulky 3.2, konkrétně vysvětlení, proč jsou v každé „buňce“ v případě Monte Carlo a naivního odhadu k dispozici tři údaje (navíc číslo 0.159 v pravém dolním rohu tabulky je zřejmě překlep). Nesrozumitelná je také například věta ze strany 28: „... X_j jsou nekorelované, proto by neměla být aplikována žádná z potenciálních transformací při výstavbě regresního modelu.“ Nerozumím také tvrzení ze spodu strany 35, kde se píše, že při použití pouze DAF3 algoritmus při naivním odhadu RMSE selhal. Jak je tudíž možné, že v tabulce 3.6 je dostupný údaj o mediánu rozdělení naivního odhadu RMSE?

Místy se objevují nepřesné, vágní až nekorektní formulace (např. tvrzení ze str. 4, že regresní analýza odpovídá na otázky, jaká je závislost mezi vysvětlující proměnou a jednou nebo více vysvětlovanými proměnnými“). Místy jsou nekonzistentně používány symboly a označení (počet vysvětlujících proměnných je označen jednou jako p , viz rovnice 1.13, následně jako k , viz vektor pod rovnicí 1.26, jindy má vektor regresních

parametrů (včetně absolutního členu) dokonce n prvků, viz začátek sekce 1.2.1, poté zase p , viz sekce 1.3.2). V praktické části na str. 30 si nejsem jist korektností rovnice 3.15: Nemá být chybová složka uvnitř exponenciální funkce (viz též rovnice 3.1 a tabulka 3.1, kde tomu tak je)?

V praktické části práce by bylo pro čtenáře také zajímavé dozvědět se i podrobnější statistiky týkající se počtu odstraněných pozorování (při aplikaci DAF1), frekvence jednotlivých transformací (při aplikaci DAF2) či frekvence počtu vysvětlujících proměnných ve finálním modelu (při aplikaci DAF3).

Prosím autora diplomové práce, aby v průběhu obhajoby zodpověděl na následující dotazy (řadím je podle důležitosti):

1. Pokud byste vyšel z obrázku 3.16b na str. 43 (výsledky pro Efekt, Hetero model, DAF3), je bootstrapový odhad RMSE možné jednoznačně hodnotit jako lepší ve srovnání s naivním odhadem? Prosím o komentář.
2. Bylo by možné alespoň kvalitativně vysvětlit vysokou hodnotu RMSE z Monte Carla pro DAF2 v tabulce 3.12 na str. 43, tj. číslo 1.161×10^{16} ?
3. V diplomové práci vycházíte částečně z článku Faraway (1992). Jaké jsou závěry v článku Faraway (1992) a jak jsou tyto závěry v souladu či v rozporu se závěry Vaší diplomové práce?
4. Vysvětlíte, jaká je motivace pro užití šikmosti rozdělení reziduí pro výběr vhodné transformace v kontextu DAF2 (viz str. 15).
5. V sekci 1.3.4 je uvedeno, že zpětná eliminační metoda může selhat. Za jakých okolností to nastává?

Nadprůměrně náročné cíle diplomové práce se podařilo splnit a student jednoznačně prokázal, že obtížnému tématu rozumí a korektně jej uchopil a zpracoval. Místy se v textu vyskytují těžkopádné formulace, nepřehledně členěný text a nepřesnosti. Velké množství (maticových) vztahů, identit, a výsledků, jež student v teoretické a praktické části uvádí, je však korektních a správně vysvětlených. Závěry z diplomové práce jsou zásadní pro praktické užití regresní analýzy. Celkový dojem z diplomové práce je pozitivní. Pokud student dokáže reagovat na dotazy položené při obhajobě, doporučuji hodnotit diplomovou práci stupněm **v ý b o r n ě**.

6. srpna 2018,

Mgr. Milan Bašta, Ph.D.

Katedra statistiky a pravděpodobnosti, FIS, VŠE, Praha