Oponentský posudek na diplomovou práci

Bc. Hana Jelínková: Porovnání prediktorů

Předložená práce se zabývá statistickými testy sloužícími ke zjištění, která ze dvou skupin prediktorů má silnější vztah k odezvě. Souhrnně lze říci, že práce zobecňuje testy o rovnosti dvou korelačních koeficientů. Text začíná jednostránkovým úvodem, ve kterém autorka seznamuje čtenáře s cíly a též obsahem práce. Následuje první (též jednostránková) kapitola, která dosti zhuštěnou formou zavádí základní značení a pojmy vesměs související s teorií testování statistických hypotéz. Vše je popsáno formou souvislého textu, ze kterého se značným způsobem vytratila matematická přesnost (viz též konkrétní připomínka č. 1 níže).

Hlavní část práce začíná druhou kapitolou, ve které jsou odvozeny tři přístupy k testování rovnosti dvou korelačních koeficientů, všechny založené na předpokladu trojrozměrného normálního rozdělení pro náhodný vektor reprezentující odezvu a dva prediktory. Jedná se o Hotellingův test, který ignoruje variabilitu prediktorů, Williamsův test a asymptotický test poměrem věrohodností. Třetí kapitola následně nabízí zobecnění na situaci, kdy srovnáváme vliv dvou skupin prediktorů na odezvu. Na základě teorie maximální věrohodnosti je zde odvozen asymptotický test. Druhá a třetí kapitola pokrývají teoretická odvození jednotlivých postupů. Zejména v případě Williamsova testu (oddíl 2.1.2) přitom autorka pouze ve velice omezené míře mohla čerpat z literatury a většina výpočtů a zdůvodnění tedy byla získána jejími vlastními úvahami. Text obou teoretických kapitol je dle mého názoru prost závažných pochybení, nicméně obdobně jako u první kapitoly mám výhrady ke zpracování. Téměř vše je psáno formou souvislého textu, kdy je často obtížné odlišit, co jsou předpoklady, co je tvrzení, resp. co odkud plyne. Například zásadní předpoklad o normálním rozdělení je učiněn mezi řečí v úvodu druhé kapitoly na str. 3 formou „Hypotézu H_0 budeme testovat ... a to na základě náhodného výběru ... z trojrozměrného normálního rozdělení ... “. Viz též konkrétní připomínka č. 3. Odvozená rozdělení jednotlivých testových statistik jsou vesměs aproximativní. Při jednotlivých odvozeních je přitom postupně použita celá řada aproximací (středních hodnot a rozptylů), resp. jsou neznámé parametry nahrazovány jejich odhady. Konečné zdůvodnění týkající se (byť aproximativního) rozdělení jednotlivých testových statistik není následně úplně vždy podloženo patřičnými teoretickými úvahami, ale spíše intuicí. Autorka se také nevyhnula jistým nepřesnostem, které jsou striktně řečeno chybami, viz konkrétní připomínky č. 4 a 5. Celkově je však třeba říci, že dvě zásadní kapitoly diplomové práce mají dobrou logickou strukturu a též se z nich dá usuzovat, že je autorka psala s porozuměním.

Práce pokračuje čtvrtou kapitolou, ve které autorka pomocí simulačních studií porovnává testy odvozené v druhé kapitole. Srovnání jsou provedena jak vzhledem k dodržování hladiny (odvozená rozdělení testových statistik jsou vesměs aproximativní), tak vzhledem k síle. Za prezentaci výsledků simulací je potřeba autorku pochválit. Výstupy jsou prezentovány poměrně přehledně a zejména je třeba vyzdvihnout fakt, že autorka pouze slovně nepopisuje tabulky čísel, ale snaží se na několika místech dát do souvislosti numerické a teoretické výsledky, což u diplomových prací na oboru PMSE nebývá bohužel pravidlem. Pouze bych chtěl varovat před snahou o přílišné zobecňování výsledků simulací, jako je tomu např. ve shrnutí na str. 30. Je potřeba mít na paměti, že simulační studie byla provedena pouze pro konečný a poměrně omezený počet scénářů.

V páté kapitole jsou odvozené postupy ilustrovány na analýze reálných dat pocházejících z Přírodovědecké fakulty UK v Praze. Aplikace řeší zajímavý a poměrně aktuální problém a to, zda prospěch na vysoké škole (PřF UK) je lépe vysvětlen prospěchem na střední škole, nebo výsledky u přijímacích zkoušek. S ohledem na předchozí teoretickou část práce se mi jako mírně nadbytečné cvičení z regrese jeví zařazení pasáží zkoumajících vliv pohlaví, nehledě na to, že např. vyřazení interakce mezi pohlavím a prospěchem na střední škole z modelu M_2^{int} je mírně problematické. Diplomová práce končí jednostránkovým závěrem, který víceméně shrnuje obsah předchozího textu.

Typograficky je práce na slušné úrovni. Vyskytuje se zde sice jisté množství překlepů (např. ,,Tučný písmem... " namísto ,,Tučným písmem... " na str. 2), nicméně s ohledem na povahu a rozsah práce je jejich počet na přijatelné úrovni.

Celkově se jedná o solidní práci, kterou **lze** nepochybně uznat jako práci diplomovou pro studijní obor Pravděpodobnost, matematická statistika a ekonometrie na Matematicko-fyzikální fakultě Univerzity Karlovy v Praze a tudíž ji **doporučuji** k obhajobě.

V Praze dne 30. srpna 2011

RNDr. Arnošt Komárek, Ph.D.
oponent diplomové práce

Konkrétní připomínky

1. Str. 2 (−14): Jednou z nejvíce matoucích věcí v první kapitole je zavedení/nezavedení pojmu *testové statistiky*, kdy není žádným způsobem uvedeno, jak tato souvisí s náhodným vektorem $\boldsymbol{X}$ reprezentujícím data. Z následujících řádků se zdá, že testovou statistikou je vždy přímo onen náhodný vektor $\boldsymbol{X}$. V průběhu práce se však za testovou statistiku berou nejrůznější funkce ,,datového" náhodného vektoru.
2. Str. 7 (−1): $\boldsymbol{Y}'\boldsymbol{Y} = \sum_{i=1}^{n} Y_i$ má být $\boldsymbol{Y}'\boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2$.
3. Str. 9 (+4): ,,a protože b^* a s^2 jsou nezávislé, ... " Nezávislost b^* a s^2 není zcela zřejmá (s^2 funkčně závisí na b^*) . V textu se autorka ani slovem nezmiňuje, odkud toto plyne.
4. U podmíněných středních hodnot se vyskytují často chybné zápisy podmínky, kdy je výrazem typu $X = x$ náhodná veličina (podmíněná střední hodnota) změněna na (nenáhodnou) reálnou funkci, z níž se však následně počítá střední hodnota, resp. jiný moment. Jde např. o výpočet rozptylu (2.11) na str. 10 nebo výpočet kovariancí na začátku i konci str. 12.
5. Str. 10: ,,... symboly s_1 a s_2 zde označují výběrové rozptyly... ". Dle použití se zdá, že s_1 a s_2 jsou spíše výběrové směrodatné odchylky a nikoliv rozptyly.
6. Str. 13 (−7): Výraz ,,rozptyl statistiky" by měl být spíše nahrazen výrazem ,,odhad rozptylu statistiky".
7. Str. 38: výraz $\lambda_{23}\, zcel_i$ ve vyjádření modelu M_2^{int} by měl být $\lambda_{23}\, ssprum_i$.

Doplňující otázky

1. Na str. 16 (uprostřed) se píše, že maximálně věrohodný odhad $\tilde{\boldsymbol{\theta}}_n$ získáme iteračně. Bylo by možné alespoň naznačit, jakým způsobem iterovat?

2. V rámci simulační studie bylo zjištěno, že Hotellingův test nedodržuje (pro uvažované scénáře) hladinu významnosti (obvykle vyšel antikonzervativně), což je způsobeno ignorováním variability prediktorů. Existují reálné situace, kdy lze doporučit použití Hotellingova testu? Jinými slovy, existují reálné situace, kdy lze ospravedlnit podmiňování realizovanými hodnotami prediktorů použité při odvození rozdělení testové statistiky?

3. V tabulce 5.1 a též dále v textu jsou uváděny též p-hodnoty jednotlivých testů. V teoretické části práce se však vždy hovoří o srovnávání hodnoty testové statistiky s kritickou hodnotou. Jakým způsobem jsou spočítány tyto p-hodnoty?

4. Vzhledem k tomu, že si nejsem zcela jist nevýznamností interakce mezi pohlavím a prospěchem na střední škole v modelu M_2^{int}, jsem zvědav, jestli by se nějak změnily závěry týkající se nevýznamného rozdílu mezi vlivem faktorů `zcel` a `ssprum` na prospěch na VŠ, jestliže by se vzalo v potaz pohlaví.