Posudek oponenta bakalářské práce

Název: Regresní analýza a spliny

Autor: Milan Benko

Shrnutí obsahu práce

Práce se zabývá využitím splinů ve statistické regresní analýze. Po úvodní první kapitole jsou ve druhé kapitole shrnuty základní poznatky z teorie lineárních modelů a související metody nejmenších čtverců. Hlavní část práce je k nalezení ve třetí kapitole. Autor začíná od po částech konstantní regresní funkce a postupně intuitivním způsobem vybuduje pojem splinu. V závěru této kapitoly jsou uvedeny též výsledky numerických studií (a) srovnávajících kvalitu odhadů při dvou různých modelech (regresní versus přirozené spliny), (b) vyhodnocujících schopnost vybraných postupů zvolit správný model. Práce je zakončena shrnutím ve čtvrté kapitole.

Celkové hodnocení práce

Téma práce. Práce zcela jistě naplňuje zadání. Při práci na tématu musel student pracovat s několika, vesměs knižními zdroji. V tomto ohledu bylo zpracovávané téma spíše náročnějším. Rozsah výsledné práce je následně spíše vyšší, patrně by neškodilo některé části vynechat. Nicméně je třeba vyzdvihnout autorovu snahu o komplexní zachycení problematiky splinů v kontextu regresní analýzy zahrnující, mimo jiného, volbu báze, volbu umístění a počtu uzlů apod. S ohledem na následnou snahu vše vměstnat do ještě stále přiměřeného počtu stran jsou však některé části až moc faktografické a prakticky u všech odvození apod. v těchto částech se odkazuje na literaturu.

Vlastní příspěvek. Student vlastními slovy a poměrně přehlednou formou shrnul značně obsáhlou problematiku. Schopnost používat popisované postupy v praxi prokázal na numerických studiích.

Matematická úroveň. Student předloženou prací projevuje schopnost formulovat matematický text. V odvozeních se dle mého názoru nevyskytují hrubé chyby. Některé výrazy v práci uvedené jsou však přinejmenším neobratné. Například:

1. Výraz (2.23) na str. 12, který je nazván jako „předpověď individuální hodnoty“ není předpovědí, ale naopak předpovídanou hodnotou.
2. „po částech spojitá polynomická funkce“ (str. 16) má být spíše „po částech polynomická spojitá funkce“.

Taktéž jsem nalezl chybu, a to ve výrazu (2.16) na str. 9 má být místo jednotkové matice uvedena hat matice $\boldsymbol{H}$. V potřebném odvození o několik řádek výše je toto však uvedeno správně. Celkově, jak píši výše, by bylo pro účely bakalářské práce vhodné vynechat faktografické pasáže a to i za cenu, že problematika splinů v regresi bude probrána v menší míře.

Jistou vadou na kráse jsou z mého pohledu též odkazy na „černé skříňky“ představované funkcemi software R. Např. na str. 5: „V praxi se však tento postup výpočtu odhadov regresných koeficientov, tj. dosadením do (2.7) a (2.8), nepoužíva. Vhodnejšou alternatívou je použitie vbudovanej funkcie `lm()` v softvéri R.“

Práce se zdroji. Student vycházel z několika vesměs knižních zdrojů. Vše je řádně citováno. Pokud mohu soudit, v práci se nevyskytují doslova přeložené pasáže a ani se nezdá, že by jednotlivé části předložené práce těsně sledovaly tu či onu knihu (jak se často u bakalářských prací stává). Drobnou připomínku bych měl ke způsobu citování ve stylu „[Číslo]“ bez uvedení jmen autorů v textu. Tento

styl značně znepříjemňuje četbu, neboť bez neustálého nahlížení do seznamu literatury nelze ani odhadovat, jaká práce je citována. V dalších odborných textech bych autorovi doporučil citace ve formě „Autor [Číslo]“ nebo ještě lépe „Autor (Rok)“.

Formální úprava. Po formální stránce je práce na dobré úrovni. Autor se nicméně nevyhnul několika překlepům či jiným nepřesnostem. Například,

1. „. . . bude uvažovať“ na str. 16 má být „. . . budeme uvažovať“.
2. Na konci strany 32 se odkazuje na obrázek 3.7. Správně by se mělo odkazovat spíše na obrázek 3.9.

Připomínky a otázky

1. Na str. 5 je výše zmíněný odkaz na funkci `lm` programu R: „V praxi se však tento postup výpočtu odhadov regresných koeficientov, tj. dosadením do (2.7) a (2.8), nepoužíva. Vhodnejšou alternatívou je použitie vbudovanej funkcie `lm()` v softvéri R.“ Mám za to, že jakýkoliv software jenom implementuje matematicky odvozené postupy a algoritmy. Jestliže funkce `lm()` nepoužívá autorem uvedené vztahy (2.7) a (2.8), jakým způsobem počítá odhady?

2. Na straně 10 se píše: „Ak nulovú hypotézu H_0 nezamietame v prospech alternatívy, znamená to, že platí podmodel $\mathbf{Y} \sim \mathsf{N}_n(\mathbf{X}_0\boldsymbol{\beta}_0, \sigma^2\mathbf{I}_n)$“. Opravdu lze tvrdit, že platí podmodel, jestliže nezamítneme příslušnou nulovou hypotézu? Jaká je správná interpretace rozhodnutí o nezamítnutí nulové hypotézy? Co z toho plyne pro použitelnost F-testu o podmodelu pro hledání vhodného modelu? Lze toto dát do souvislosti s vašimi zjištěními v odstavci „F-test“ na str. 35?

3. Na str. 32 jsou uvedena některá nastavení pro numerickou ilustraci. Jak rozsahy výběru (512, resp. 32), tak směrodatné odchylky chyb (0.108, resp. 0.579) jsou voleny jako ne zcela „hezká čísla“. V obdobných situacích bývá zvykem volit tyto hodnoty spíše tak, že je zadána pouze cifra odpovídající řádu zvolené hodnoty (např. místo 512 by bylo voleno 500). Existuje nějaká motivace pro hodnoty zvolené autorem? Taktéž bývá slušností uvést vše potřebné pro možnost vše přepočítat, což není jenom uvedení, že byla použita funkce `set.seed`. V autorově kontextu chybí počet a umístění uzlů, resp. regresní koeficienty u „skutečné“ regresní funkce.

Závěr

Práci považuji za 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 31. srpna 2015