Posudek vedoucího na bakalářskou práci

Autor bakalářské práce: Milan Benko

Název bakalářské práce: Regresní analýza a spliny

Vedoucí bakalářské práce: Mgr. Milan Bašta, Ph.D.

Cílem předložené bakalářské práce je představit modely regresních splinů (s jednou vysvětlující proměnnou) jako potenciálně vhodné modely v kontextu regresní analýzy a studovat jejich vlastnosti.

Obsah práce

V kapitole 2 se autor zabývá úvodem do lineárních regresních modelů (regresním modelem přímky, modelem vícenásobné lineární regrese, odhadem metodou nejmenších čtverců a jeho vlastnostmi, podmodelem, odhadem střední hodnoty vysvětlované proměnné v novém bodě, kritérii pro výběr regresního modelu).

V kapitole 3 jsou již představeny modely regresních splinů. Autor prezentuje dvě možné báze pro reprezentaci regresních splinů, a to useknutou mocninnou bázi a B-spliny. Zabývá se také krátce volbou polohy a počtu uzlových bodů v modelu regresního splinu. Následně představuje pojem přirozeného kubického splinu, pro který odvozuje vhodnou bázi pro jeho reprezentaci. Kvalitativně také porovnává střední čtvercovou chybu odhadu regresní funkce v modelu regresního a přirozeného splinu při platnosti modelu regresního splinu. Pomocí Monte Carlo simulací následně v kontextu modelů splinů ilustruje, že běžná kritéria pro výběr regresního modelu, jež poskytuje přesnější odhad regresní funkce, nefungují "bezchybně".

Bakalářská práce byla vypracována v souladu s předloženými zásadami pro vypracování. Práci by prospělo, kdyby bylo o něco méně prostoru věnováno úvodu do lineární regrese (kapitola 2) ve prospěch pasáží zabývajících se modely splinů (kapitola 3). V těchto pasážích mohly být ještě detailněji rozvedeny např. výhody modelů regresních splinů oproti modelům regresních polynomů, výhody a nevýhody B-spline báze oproti mocninné useknuté bázi. O malinko podrobnější mohla býti také diskuze a případná analýza týkající se volby počtu a polohy uzlových bodů.

Vlastní přínos autora

V práci shledávám několik vlastních příspěvků autora a to: odvození báze pro reprezentaci přirozených kubických splinů (str. 27 až 29), kvalitativní diskuzi konceptu "bias-variance trade-off" v kontextu přirozených a regresních splinů (str. 30 až 32) a naprogramování řady vlastních funkcí v softwaru R, které autor následně využil k ilustraci použití modelů splinů na reálných datech a také v kontextu Monte Carlo simulací.

Korektnost formulací a matematického textu

V textu práce se vyskytují místy nekorektní formulace. Jako příklad je možné uvést kapitolu 2.5.1, ve které autor poněkud nekorektně používá pojmy "hodnota vysvětlované proměnné v novém bodě", "předpověď (individuální) hodnoty vysvětlované proměnné v novém bodě" a "odhad střední hodnoty vysvětlované proměnné v novém bodě" (viz např. rovnice 2.23 na str. 12, ve které autor "hodnotu vysvětlované proměnné v novém bodě" nekorektně nazývá "předpovědí", nebo i dále v textu). V Závěru na str. 37 autor uvádí, že v rámci Monte Carlo simulací zkoumal, zda kritéria pro výběr modelu dokáží správně rozhodnout ve prospěch skutečně platného modelu. Toto však není korektní formulace, protože autor v Monte Carlo simulacích studoval, zda kritéria pro výběr modelu dokáží identifikovat model lepší z hlediska přesnosti odhadu regresní funkce – tento model však nemusí být vždy skutečně platný. V textu se vyskytují místy i další nekorektní či mírně matoucí formulace.

V textu je také relativně nemálo těžkopádných formulací (viz např. str. 21: "Báza funkcie spojitého polynómu v K uzloch stupňa d ...", ale i jinde), místy až nesrozumitelných (viz např. vysvětlení ihned pod rovnicí 3.38 na str. 32, které se autorovi nepodařilo srozumitelně napsat, ač lze i přesto vytušit, co jím chtěl autor říci, nebo str. 32: "...a tú si následne vyhodnotíme v nami uvažovanom maximálnom počte hodnot nezávislých premenných." – jaký maximální počet má autor na mysli?).

V matematickém zápise se místy vyskytují překlepy a chyby (např. v rovnici 3.4 na str. 18 má být dolní index $\mathcal{I}_K$ a nikoliv $\mathcal{I}_k$; hypotéza pod rovnicí 3.5 na str. 19 je formulovaná nekorektně; dolní index k nemůže pod rovnicí 3.10 na str. 22 nabývat hodnoty $K + d + 1$, jak autor tvrdí; matici $\mathbf{X}$ v rovnici 2.26 na str. 12 zapomněl autor uvést tučně; matice $\mathbf{X}^T\mathbf{X}$ na str. 7 dole měla být v kontextu okolního textu lépe označena jako pozitivně definitní a nikoliv jako pozitivně semidefinitní; spojení definice pozitivní definitnosti matice a definice symetrické matice ve společné definici 3 na str. 8 není vhodné; nerozumím také tomu, proč se autor v důkazu tvrzení 1 na str. 8 zabývá symetrií matice $\mathbf{X}^T\mathbf{X}$; apod.). Některá značení nejsou konzistentní (např. na str. 12 nahoře označuje autor prvky vektoru $\mathbf{x}_*$ nejprve netučně, následně však tučně; v tabulce 3.1 je počet pozorování značen jako n_j, v tabulce 3.2 a 3.3 jako $\mathbf{n}$ apod.).

Další připomínky a postřehy

Nastavení Monte Carlo simulací je v některých pasážích nejasné. Bylo by např. vhodné podrobněji popsat, jak vypadá regresní funkce, která je v simulaci uvažována (pokud je neměnná, bylo by vhodné ji vykreslit apod.). Závěry z Monte Carlo simulací jsou možná až příliš stručné, trochu podrobnější diskuze výsledků by byla vhodnější.

Práce s literaturou je na relativně dobré úrovni, doporučil bych však řazení položek v seznamu literatury podle abecedy.

Místy je v práci zmiňována implementace postupů v softwaru R. Někdy však tato zmínka není umístěna na vhodné místo v textu a narušuje jeho plynulost (viz např. začátek sekce 2.5.3 na str. 13).

Celkové hodnocení

V bakalářské práci jsou pěkné pasáže, viz např. autorův vlastní přínos zmiňovaný výše. Některé zajímavé věci, které by v práci mohly být ještě uvedeny, v ní však scházejí. Práce trpí místy nekorektními formulacemi a značením, vyskytují se v ní místy těžkopádná a nesrozumitelná vyjádření. Na kvalitě práce ubírají také občasné překlepy ve slovech či nedokončené věty (viz např. na str. 11 dole "Tejto téme ako aj", ale i jinde). Práci Milana Benka přes výše uvedené výhrady doporučuji uznat jako bakalářskou práci.

Otázky

1. Vysvětlete rozdíl mezi odhadem střední hodnoty vysvětlované proměnné v novém bodě a předpovědí individuální hodnoty vysvětlované proměnné v novém bodě.
2. Pro jaká data v praxi byste doporučil použití splinů a pro jaká nikoliv?

25. srpna 2015

Mgr. Milan Bašta, Ph.D.,
Katedra statistiky a pravděpodobnosti,
FIS, VŠE, Praha