Posudek oponenta bakalářské práce

Název: Omezené a cenzorované vysvětlované proměnné
Autor: Rudolf Kostka

Shrnutí obsahu práce

Práce si klade za cíl pojednat o regresních modelech s dichotomickou a multinomickou odezvou, respektive o modelech, kde je odezva vystavena cenzorování. První kapitolu lze charakterizovat jako úvod do problematiky zobecněných lineárních modelů. Druhá kapitola, která vesměs vznikla překladem řešených ilustračních příkladů z knihy D. Colletta, ukazuje použití vybraných modelů na reálných datech. Třetí kapitola je potom věnována problematice modelování cenzorovaných dat. V rámci třetí kapitoly též autor provádí analýzu dat, která pro účely bakalářské práce získal od firmy SC Johnson. Poslední, čtvrtá kapitola potom srovnává tři programové balíky, které lze použít k odhadů některých modelů používaných v analýze přežití.

Celkové hodnocení práce

Celkově se jedná o velice špatnou práci. Matematické části textu jsou psány dosti povrchně, s chybami a působí na mě, jako kdyby byly nahodile poskládány za sebe bez bližšího porozumění. Části textu věnované praktické analýze pomocí statistických programových balíků se zase omezují na tisk kódů a výstupů, které jsou jenom velice stručně komentovány, opět bez alespoň trochu hlubšího vhledu do zkoumané problematiky.

Téma práce. Téma se mi zdá poměrně obsáhlé pro bakalářskou práci. Již samotný název vede na dvě rozsáhlé třídy modelů: (i) s omezenou odezvou, (ii) s cenzorovanou odezvou. Obě tyto třídy jsou samy o sobě dosti obsáhlé pro bakalářskou práci, není-li přesněji specifikováno, čím se práce má zabývat. Výsledkem je potom dosti povrchní zpracování, dle mého názoru povrchnější než je vyžadováno od bakalářské práce na studijním programu Matematika.

Vlastní příspěvek. Za vlastní příspěvek autora lze snad považovat fakt, že autor *použil* tři různé programové balíky na reálná data firmy SC Johnson a stručně popsal získané výstupy. Nicméně je nutné podotknout, že se skutečně jedná o pouhý popis výstupů s naprosto minimální přidanou hodnotou. Příslušné pasáže textu tedy rozhodně nelze nazývat zprávou o provedené statistické analýze.

Matematická úroveň. Matematická úroveň práce je podprůměrná až nevyhovující. Matematický text v práci obsažený nelze v žádném případě považovat za rigorózní a korektně zformulovaný. Kromě toho, že se v práci téměř nevyskytují řádná odvození alespoň některých uváděných vztahů, považuji za jeden z významných nedostatků fakt, že se v textu objevují výrazy obsahující symboly, které nejsou nikde definovány (v tom horším případě), resp. jsou definovány až o několik stran později. Nabývám tudíž dojmu, že autor ne vždy ví, s jakými matematickými objekty vlastně pracuje. První (a zdaleka ne poslední) nedostatek tohoto typu nalézám hned v prvním „matematickém“ výrazu na začátku první kapitoly, kde se vyskytuje výraz $y_t = \boldsymbol{x}_{t\cdot}\boldsymbol{\beta} + \varepsilon_t$. V blízkém textu je vágně popsáno, co je y_t, co by mohlo být $\boldsymbol{x}_{t\cdot}$. Zůstává mi však

utajeno, co má označovat symbol ε_t. Taktéž není úplně jasné, co indexuje t. Z výrazu výše si domýšlím, že jednotlivá pozorování, z výrazu $\sum_{t=1}^{p} x_t\beta_t$, který se zde vyskytuje též, bych však usuzoval na indexování jednotlivých regresorů. Konečně si u výše uvedeného výrazu nejsem jist, zda si autor uvědomuje, že násobí vektory. Viz chybějící transpozice ve výrazu „$\boldsymbol{x}_t.\boldsymbol{\beta}$“, která se v práci opakovaně vyskytuje.

Obecně matematické vyjadřování autora je poměrně vzdáleno rigoróznosti, kterou bych očekával v psaném textu, viz např. výraz „*rovná čára*“ uprostřed str. 10, který má patrně reprezentovat přímku.

V neposlední řadě se v práci vyskytuje nekonzistentní značení, kdy např. logaritmická věrohodnost je značena jako L v kapitole 1 a jako l v kapitole 3. Symbol l přitom v 1. kapitole označuje věrohodnost, tedy l ve 3. kapitole je logaritmem l z 1. kapitoly.

Práce se zdroji. Zdroje citovány jsou. Práce nicméně obsahuje doslova přeložené pasáže (zejména v kapitole 2). Kupříkladu text od poloviny str. 11 po začátek str. 13 lze v anglické verzi nalézt na str. 55–56 Colletovy knihy ze seznamu literatury. Obdobně text od str. 13 až po str. 16 nalezneme ve stejné knize postupně na str. 106–108, 105–106 a 110–111. Autor sice na začátku druhé kapitoly uvádí, že je tato převážně převzata z uvedené knihy, nicméně nejsem si jist, zda je vhodné, aby několik stran bakalářské práce bylo doslovným (a ještě ne příliš zdařilým) překladem jiného zdroje.

Formální úprava. Po formální stránce je práce průměrná. Členění na kapitoly, podkapitoly atp. je v pořádku, nicméně stylisticky se jedná o spíše podprůměrnou práci. Z mnohých vět v češtině je více než zřejmé, že vznikly překladem z angličtiny (v některých případech skoro nabývám dojmu, že se jedná o nijak nekorigovaný překlad získaný některým z internetovských překladačů). Namátkou lze uvést věty: „*Poslední důvod je jaksi technický, ale lze ukázat, že binární data mohou být sumarizovány ve smyslu kvantity, známý jako postačující statistika (sufficient statistics), pokud je použita logistická transformace.*“ na str. 13 nebo „*Při odhadování veličin jako* ED_{50} *a podobně nás může zajímat standartní odchylka a interval spolehlivosti, aby náš odhad mohl být posouzen.*“ na str. 15. Taktéž překlepy lze v práci nalézt v míře větší než malé.

Ne zcela v pořádku je též sazba matematických výrazů (např. proměnná y jednou napsaná kurzívou a jednou patkovým písmem v úvodu kap. 3 nebo nevhodně zvolené velikosti závorek, viz např. str. 13, 14 a další). Není mi též jasné, proč je od oddílu 4.3 náhle vše sázeno kurzívou.

Jedním z dalších nedostatků je způsob odkazování na obrázky. Např. na obrázek 2.2 na str. 15 je odkázáno nevyhovujícím způsobem jako: „*. . . k vysvětlující proměnné dávka a log(dávka) je ukázán (2.2). (2.2) jednoznačně naznačuje, že. . .*“.

V neposlední řadě je třeba práci vytknout seznam literatury na str. 36, ve kterém jsou jednotlivé položky nekonzistentně formátovány: křestní jména autorů někde v plné verzi, jinde pouze iniciály, příjmení jednou kapitálkami, podruhé nikoliv.

Připomínky

Práce obsahuje nemalé množství nepřesností a faktických chyb. Některé z nich zmiňuji výše, výběr z dalších následuje:

1. Co označuje vektor $\boldsymbol{X}$ na třetím řádku str. 4?

2. Poslední vysazený matematický výraz na str. 5 není *odds ratio*, jak se autor mylně domnívá.

3. Uprostřed str. 7 nesouhlasím s tvrzením, že nějaký jev nastane, jestliže jeho odhadnutá pravděpodobnost překročí hodnotu 0,5.

4. Na str. 12 (5. řádek odspodu) klade autor rovnítko mezi tři odlišné veličiny, viz „*. . . závislost pravděpodobnosti úspěchu (též nazýváno šance – odds ratio). . .* “.

5. Anglický výraz *deviance* se v uvedeném kontextu rozhodně nepřekládá jako *odchylka* (viz např. str. 15).

6. Nesouhlasím s pasáží: „*. . . pouze víme, že něčí plat přesahuje hranici 100 000, ale již nevíme o kolik. V takovém případě všechny tyto platy budeme považovat za rovny 100 000 a mluvíme o cenzorované vysvětlované proměnné.*“ Je-li nějaká veličina cenzorovaná, nepovažuje se za rovnu hodnotě, ve které je cenzorována.

7. Uprostřed str. 21 autor tvrdí, že „$\lambda_0(\tau)$ *je bazická intenzita úmrtnosti a není závislá na čase* t.“ Ze zápisu používaného v bakalářské práci, kdy písmenem t se označuje tu čas, jinde indexy jednotlivých pozorování se skutečně může zdát, že λ_0 nezávisí na čase. Ale je tomu skutečně tak?

8. Kaplan a Meier navrhli *odhad* funkce přežití, nikoliv *model*, jak autor tvrdí v horní části str. 22.

9. Výraz na str. 22 uvedený slovy „*Podmíněná pravděpodobnost pak vypadá*“ neudává pravděpodobnost, ale její *odhad*.

10. V poznámce 4.2 na str. 26 autor zmiňuje jakousi proměnnou „cens“. Není jasné, co se tímto myslí.

11. Není vysvětlen význam přerušovaných čar na obrázcích 4.2 a 4.4.

Otázky

1. Na začátku str. 7 tvrdíte, že se pomocí McFaddenova koeficientu posuzuje kvalita odhadů regresních parametrů. Mohl byste vysvětlit, v jakém smyslu máte definovánu onu kvalitu odhadů?

2. Na str. 15 tvrdíte, že při odhadování jistých veličin „*nás může zajímat též standartní odchylka a interval spolehlivosti, aby náš odhad mohl být posouzen.*“ Jakým způsobem se pomocí směrodatné odchylky, resp. intervalu spolehlivosti *posuzuje* statistický odhad?

3. Byl byste schopen matematicky zdůvodnit, proč je v případě useknuté odezvy v logaritmické věrohodnosti na konci str. 19 zahrnut faktor
$$-\ln\left\{F\Big(\frac{h_t-\boldsymbol{x}_{t\cdot}^\top\boldsymbol{\beta}}{\sigma}\Big)-F\Big(\frac{d_t-\boldsymbol{x}_{t\cdot}^\top\boldsymbol{\beta}}{\sigma}\Big)\right\}?$$

4. Na konci str. 21 v kontextu Coxova modelu píšete, že „*bazická intenzita úmrtnosti a parametry* $\boldsymbol{\beta}$ *se odhadnou metodou maximální věrohodnosti.*“ Můžete napsat věrohodnost tohoto modelu a nastínit, jak ji maximalizovat? Není pro maximální věrodhodnost potřeba specifikovat S_0, resp. λ_0?

5. Na začátku textu na str. 24 píšete, že „*Vidíme, že P-hodnota je u modelu menší než 0.05. U zbývajících modelů je tomu také tak a tak jsou všechny modely přípustné.*“ Mohl byste specifikovat, k jakému testu se váže zmiňovaná P-hodnota a jak tento test souvisí s „přípustností“ modelu? Co myslíte touto přípustností?

Závěr

Práci považuji za **nevyhovující** a **nedoporučuji** 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 16. června 2014