Oponentský posudek na doktorskou dizertační práci

Mgr. Petr Klášterecký: Odhady parametrů v subkohortních studiích

Autor se ve své práci věnuje odhadu parametrů v regresních modelech pro analýzu zprava cenzorovaných dat při použití subkohortních (*case-cohort*) studií. Jedná se o techniky, jež mají za hlavní cíl snížit rozsah výběru (při zachování předepsané síly testů, či přesnosti odhadů) při analýze času do nejaké události v situacích, kdy se sledovaná událost vyskytuje s poměrně malou pravděpodobností. Vzhledem k tomu, že modely tohoto typu umožňují výrazně snížit náklady na provedení takové studie, jedná se v dnešní době o dosti aktuální téma.

Práce je rozdelněna do šesti kapitol. Po krátké úvodní kapitole následuje kapitola 2, v níž je zavedena hlavní část značení a jsou popsány některé regresní modely používané v analýze zprava cenzorovaných dat. Vedle nejběžnějšího Coxova modelu proporcionálního rizika se autor venuje též méně běžnému modelu proporcionálních šancí a nelineárním transformačním modelům. Pro všechny uvedené modely jsou diskutovány metody odhadu při použití běžné (kohortové studie). V závěru druhé kapitoly je představen též model logistické regrese, který později (v kapitole 4) poslouží jako stavební kámen pri odvození nové metody pro analýzu subkohortních studií. Kapitola 3 nejprve zavádí pojem subkohortní (*case-cohort*) studie a diskutuje přístupy známé z literatury k odhadu regresních modelu zavedených v kapitole 2 při použití tohoto druhu studie. Prezentované přístupy využívají vesmés vážení založené na pravděpodobnosti výběru do subkohorty. Součástí třetí kapitoly je též simulační studie, která ukazuje, že přístupy známé z literatury selhávají a to zejména v situaci nízkých pravděpodobností výběru do subkohory a korelovaných vysvětlujících proměnných. Oba tyto jevy jsou přitom poměrne časté v praktických, zejména pak epidemiologických aplikacích. Simulační část kapitoly 3, autorem publikovaná též ve sborníku odborné konference, pěkným zpusobem motivuje další část dizertační práce.

Hlavní teoretická část dizertační práce je obsažena v kapitole 4, v které autor chytre využívá souvislostí mezi logistickou regresí a modelem proporcionálních šancí a odvozuje novou metodu odhadu regresních parametru v subkohortních studiích bez nutnosti používat váhy založené na pravděpodobnostech výběru do subkohorty. V části 4.2 jsou odhad a jeho vlastnosti nejprve odvozeny neformálně, v části 4.3 jsou potom tyto vlastnosti zformulovány do vět a formálně dokázány při předpokladu konstantního cenzorování a časově konstantních vysvětlujících proměnných. Následující 5. kapitola ilustruje na simulační studii chování autorem navržených odhadu a srovnává je s odhady dríve publikovanými v literatuře. Práci zakončuje shrnutí a diskuze v 6. kapitole.

Obecnější připomínky, témata k diskuzi

1. Téma práce je poměrně jasně motivováno praktickými potřebami a zejména pak snahou snížit finanční náročnost studií zabývajících se řídce se vyskytujícími jevy. Překvapuje mě proto, že celá práce neobsahuje vubec žádnou aplikaci autorem vyvinutých metod na reálná data. Reálný ilustrační příklad, který by čtenáře provázel celou prací, by nejen více motivoval čtenáře, ale v mnoha ohledech by též výrazně zlepšil čtivost a atraktivitu celé práce.

2. V celé práci jsou uvažovány pouze časově konstantní vysvětlující proměnné, což je pro praktické aplikace poměrně omezující podmínka. Též vysvětlující proměnné zmíněné v motivačním

úvodu kapitoly 3 na str. 15 (krevní vzorky, laboratorní testy) jsou spíše závislé na čase než naopak. Domnívám se, že s ohledem na důležitost tohoto zobecnění, by mu mělo být věnováno trochu více místa než tomu je v předložené práci (poznámka pod čarou na str. 16, krátký odstavec v závěru na str. 77).

3. Přehlednost práce do jisté míry trpí zjevným autorovým odporem ke kreslení obrázku (v celé práci se vyskytují pouze dva na str. 30). Kromě níže zmíněného grafického vysvětlení principu subkohortní studie by též výsledky simulačních studií byly dle mého názoru přehlednější, jestliže by některá čísla z mnoha tabulek byla nahrazena vhodnými obrázky (např. krabičkovými grafy odhadu spočtených na jednotlivých simulovaných datech).

4. Pro účely simulačních studií v kapitole 5 autor nepochybně softwarově implementoval metody navržené a studované v kapitole 4. Jsou tyto implementace řádně zdokumentovány tak, aby jich někdo jiný mohl použít k analýze vlastních dat? Jsou tyto implementace dostupné široké odborné veřejnosti?

5. Výsledky prezentované zejména ve čtvrté kapitole, spolu s případnou softwarovou implementací (o jejíž existenci nepochybuji) a aplikací na reálná data (jež by byla žádoucí) jsou nepochybně velmi dobrým základem pro publikaci v kvalitním odborném časopise. Má autor představu o titulu časopisu, v kterém by své výsledky chtěl publikovat a v jakém časovém horizontu?

Konkrétní připomínky

1. „At risk" proces $Y(t)$ na str. 5 by měl být definován jako $Y(t) = I[X \geq t]$ a nikoliv jako $Y(t) = I[X > t]$ jak je uvedeno v práci.

2. Prosím o vysvětlení posledního odstavce na str. 8. Tvrdí se zde, že situace, kdy $S(t|Z = 1) \approx S(t|Z = 0) \approx 0$ odpovídá nízkým pravděpodobnostem události (řídkým jevům). U řídkého jevu (s malou pravděpodobností události) bych však spíše očekával $S(t|Z = 1) \approx S(t|Z = 0) \approx 1$. Myslím, že je potřeba při definici šance zaměnit význam „úspěchu" a „neúspěchu", aby bylo možné ukázat vztah mezi Coxovým modelem a modelem proporcionálních šancí.

3. Vysvětlení principu subkohortní studie (hlavní téma celé práce) ve 2. odstavci na str. 16 mi při prvním čtení nebylo příliš jasné. Čtenář je navíc mírně zmaten poslední větou tohoto odstavce, jež se vztahuje k *nested case-control* studii, zmíněné v jednom z předcházejících odstavců. Na tomto místě by bylo vhodné umístit názorný diagram (ukazující odlišnosti subkohortní studie od dříve zmíněných typů studií) či vysvětlit principy na konkrétním příkladě.

4. Pro získání optimálního logistického odhadu (str. 14 a dále) je potřeba podstoupit optimalizaci při omezeních $\sum_{k=1}^{K} w_{j,k} = 1$, $0 \leq w_{j,k} \leq 1$, $j = 1, \ldots, p$. Neuvažoval autor o vhodné transformaci vah, která by odstranila nutnost vázané optimalizace? V podobných situacích se často používá následující transformace ($j = 1, \ldots, p$):

$$w_{j,k} = \frac{\exp(a_{j,k})}{\sum_{l=1}^{K} \exp(a_{j,l})}, \qquad k = 1, \ldots, K,$$
$$a_{j,1} = 0,$$

a optimalizace vzhledem k $(a_{j,2}, \ldots, a_{j,K})' \in \mathbb{R}^{K-1}$, $j = 1, \ldots, p$. Váhy jsou sice potom ostře odražený od nuly a jedničky, ale domnívám se, že by to nemělo představovat výraznější problém. Bylo by možné využít tohoto přístupu v kontextu autorovy práce?

5. Lze nějak motivovat volbu hodnoty 30 ve vyjádření prahové hodnoty k_0 v horní části str. 46?

6. V kapitole 5 (simulační studie) je uvedeno, že „data" byla generována při modelech proporcionálního rizika, či proporcionálních šancí. Chybí zde však informace, jak bylo voleno základní rozdělení (pro $\boldsymbol{Z} = \mathbf{0}$). Nedostupnost této informace znemožňuje případnému zájemci zopakovat autorem prezentovanou simulační studii.

7. Data pro simulační studii byla generována při platnosti předpokládaného modelu proporcionálních šancí a při platnosti Coxova modelu proporcionálního rizika, který je při malé pravděpodobnosti události téměř ekvivalentní modelu proporcionálních šancí. Má autor nějakou představu o tom, jak se jeho odhady chovají v případě, že se model generující data výrazněji liší od předpokládaného modelu proporcionálních šancí?

Shrnutí

Autor v rámci své dizertační práce odvodil a vyšetřil vlastnosti nového odhadu regresních parametrů statistického modelu pro analýzu zprava cenzorovaných dat pocházejících ze subkohortních studií. Autorova metoda založená na váženém odhadu z posloupnosti logistických regresí je zcela novým přístupem v daném kontextu. Velice důležitým přínosem je fakt, že odvození asymptotických vlastností autorova odhadu nevyžaduje v praxi obtížně splnitelný předpoklad nenulových pravděpodobností výběru do subkohorty. V budoucnosti lze jistě očekávat publikaci dalších odborných textů, jež budou dále zobecňovat autorovu metodu, a to nejen směry popsanými v šesté kapitole. Pro další využití v praktických aplikacích považuji za velice důležité podrobnější rozpracování problému časově závislých vysvětlujících proměnných. Aktuálnost tématu dále podtrhuje fakt, že hlavní konkurenční přístupy využívající váhy založené na pravděpodobnosti výběru do subkohorty, byly publikovány vesměs v tomto století (např. Chen, 2001; Kong, Cai, Sen, 2004; Lu a Tsiatis, 2006) a v průběhu autorova doktorského studia.

Při návrhu vlastního odhadu prokázal autor nemalou invenci a schopnost tvůrčího myšlení, při odvození jeho vlastností musel využít pokročilých znalostí teorie maximální věrohodnosti a asymptotické statistiky. Pro praktický výpočet navržených odhadů bylo nutné využít též poznatků z numerické matematiky a v neposlední řadě technickou programátorskou zručnost.

Práce je po formální stránce kvalitně zpracována a s ohledem na její rozsah se v ní vyskytuje pouze zanedbatelné množství překlepů, jazykových nepřesností a typografických chyb (na několika místech je např. používána jak desetinná tečka, tak čárka).

Soudím proto, že autor prokázal schopnost samostatné tvořivé práce a jeho práci doporučuji přijmout k obhajobě.

V Praze dne 15. října 2009