

OPONENTSKÝ POSUDEK NA DIPLOMOVOU PRÁCI

Název: Modely s kategoriální odezvou

Autor: Anežka Faltýnková

Shrnutí:

Diplomová práce Anežky Faltýnkové se zabývá regresními modely pro kategoriální data, zejména logistickou regresí pro binární odezvu, multinomiální logistickou regresí pro nominální kategoriální odezvu a modelem proporcionálních šancí s kumulativními logity pro ordinální kategoriální odezvu.

Práce je kompilačního charakteru. Téma je zajímavé, svým objemem a náročností je vhodné pro diplomovou práci na oboru FPM, který Anežka Faltýnková studuje. Celkové pojetí práce, to jest střídání teoretického výkladu s praktickými a numerickými příklady doplněné jednou obsáhlejší aplikací na závěr, je velmi vhodné.

Zpracování tématu diplomantkou však vykazuje příliš mnoho závažných nedostatků. Po stránce formální září ohromující počet jazykových a gramatických chyb, překlepů, chybějícího a nekonzistentního značení, a dále i nejasné, chybějící a nedostatečně specifické citace. Matematickou stránku kazí řada fakticky chybných tvrzení, nejasné předpoklady, chybějící definice a chyby v důkazech. Po stránce koncepční lze vytknout chaotickou organizaci výkladu bez logické návaznosti, nedostatečné vysvětlení účelu některých částí, vynechání některých zásadních poznatků a souvislostí i strojně mechanické pojetí praktických příkladů bez hlubšího porozumění. Výběr nejzásadnějších konkrétních připomínek k prvním dvěma kapitolám je zařazen na konci posudku.

Předložená práce Anežky Faltýnkové podle mého názoru zdaleka nesplňuje požadavky kladené na diplomovou práci na oboru Finanční a pojistná matematika. Celkově ji hodnotím jako nedostatečnou a *nedoporučuji ji uznat za diplomovou práci.*

Vybrané zásadní připomínky: [Z důvodu úspory místa uvádím pouze připomínky ke kap. 1 a 2.]

- Kapitola 1.1 Teorie maximální věrohodnosti
 - Definice 1 obsahuje nedefinované symboly.
 - Chybí definice maximálně věrohodného odhadu.
 - Definice 7 platí pouze pro jedno určité θ , nelze ji použít jako obecnou definici Fisherovy informační matice jakožto funkce θ .
 - Věta 8 obecně neplatí. Chybí specifikace podmínek regularity.
 - Věta 9: konsistence je nutno prokázat, nikoli brát jako předpoklad.
 - Chybí věty o testování za přítomnosti rušivých parametrů. Ty jsou zásadní pro kapitolu 2.5 a 3.
- Kapitola 1.2 Odhad parametrů v modelu logistické regrese
 - Chybí vyjádření pozorované a očekávané informační matice, chybí pojednání o existenci a jednoznačnosti řešení věrohodnostních rovnic.
- Kapitola 1.3 Dummy proměnné
 - (1.6) není žádný vztah
 - Vysvětlení dummy proměnných je nedostatečné, věty nedávají smysl.
 - Ve výsledném vzorci se ztratil význam konstanty p (počet parametrů).
- Kapitola 1.4 Příklad
 - Jaká je správná interpretace parametrů logistického modelu? Toto mělo být uvedeno již v úvodu kapitoly 1.
 - Parametry jsou konstanty, tudíž jejich směrodatné odchylky jsou nulové.
- Kapitola 2.1 Logitové modely pro nominální odezvu
 - Z nemá multinomické rozdělení, speciálně nemá to konkrétní multinomické rozdělení, které autorka uvádí. Reprezentace Z pomocí Y nesouhlasí.
 - Chybí interpretace parametrů modelu (viz např. Agresti).
 - Chybí definice odhadu a pojednání o jeho vlastnostech, výpočet skórové statistiky, vyjádření informační matice. Výklad modelu nelze odbýt tím, že se uvede věrohodnostní funkce.

- Kapitola 2.2 Logitové modely pro nominální odezvu – příklad
 - Uved'te příklad do souvislosti se značením zavedeném v předchozí kapitole. Dejte konkrétní význam všem důležitým symbolům. Interpretujte odhady parametrů (jejich číselné hodnoty, nejen znaménka).
- Kapitola 2.3 Kumulativní logitové modely s ordinální odezvou
 - Co je to Z a jaké je její rozdělení? Co znamená „ Z padne pod určitý bod“?
 - Proč parametry nezávisí na j ? Je to nutnost, předpoklad, volba, kterou jste učinila?
 - Chybí interpretace parametrů modelu (viz např. Agresti).
 - Chybí definice odhadu a pojednání o jeho vlastnostech, výpočet skórové statistiky, vyjádření informační matice.
- Kapitola 2.4 Kumulativní logitové modely s ordinální odezvou – příklad
 - Neumíte interpretovat parametry modelu. Místo toho si vytvoříte jakousi tabulku procent, ze které popisujete, co se děje v datech. Na co pak máte ten kumulativní logitový model?
- Kapitola 2.5 Hledání vhodného modelu
 - Je nutno jasně definovat devianci a konkrétně uvést tvrzení, na nichž je založeno testování submodelů (včetně předpokladů).
 - Jen tato kapitola sama obsahuje nepřiměřeně mnoho nepřesných, nepodložených či nepravdivých tvrzení.
- Kapitoly 2.6 a 2.7 Hledání vhodného modelu – příklady 1 a 2
 - Uved'te u každého uvažovaného modelu počet parametrů a devianci. Ukažte, jak se počítá testová statistika a uved'te ji. Nestačí uvést pouze p -hodnotu.
 - Postup hledání vhodného modelu je chybný. Nelze pouze porovnávat deviance a nedívat se na počet parametrů. Kdyby měla rasa tři možné hodnoty místo dvou, postup by selhal.
 - Tabulka 2.6 neuvádí koeficienty, jak tvrdí legenda, ale nevysvětlené symboly.
- Kapitola 2.8 Lineární regrese versus multinomická regrese s ordinální odezvou
 - Jak byly zvoleny dělicí body pro multinomiální odezvu? Kolik pozorování se nachází v jednotlivých kategoriích? Zde je nutno postupovat velmi opatrně.
 - Koeficienty v tabulce 2.8 naznačují, že výpočet maximálně věrohodného odhadu nemusel zkonvergovat. To mohlo být způsobeno buď nevhodnou volbou kategorií odezvy nebo nevhodnou formou regresorů.
- Kapitola 2.9 Výpočetní aspekty iterativního hledání odhadů parametrů
 - Z tabulky 2.9 se zdá, že celá procedura nekonverguje.
 - Jaké kritérium používají tyto algoritmy k zastavení iterací?
- Kapitoly 2.10 a 2.11 Grafické a naivní hledání odhadů parametrů
 - K čemu jsou tyto kapitoly dobré? Proč je potřeba hledat alternativy ke standardnímu numerickému algoritmu?
 - Ani jeden ani druhý postup nefunguje, jak je koneckonců v práci přiznáno. Proč jsou tedy tyto kapitoly v práci ponechány?
- Kapitola 2.12 Generování dat pro multinomickou regresi
 - Kapitola v tištěné práci nemůže generovat náhodná data. Z nemá multinomické rozdělení. Na kolika simulacích jsou založeny tabulky 2.16 a 2.17? Vzorce (2.15) až (2.22) pouze řeší soustavu lineárních rovnic o 4 neznámých. K čemu je to dobré?

Intenzita připomínek ke zbytku práce by byla zhruba stejná jak u prvních dvou kapitol.

doc. Mgr. Michal Kulich, PhD.
 KPMS MFF UK
 31. srpna 2015