Hodnocení disertační práce v oboru Statistika (VŠE v Praze)

Jméno doktoranda: **Ing. Lukáš Sobíšek**

Název práce: Shluková a regresní analýza mikropanelových dat

Jméno hodnotitele: doc. RNDr. Arnošt Komárek, Ph.D.

Posudek

Předložená práce pojednávající o shlukové a regresní analýze mikropanelových dat je členěna do osmi kapitol a obvyklého úvodu a závěru. V první kapitole je zaveden pojem panelových (resp. longitudinálních) dat. Druhá kapitola popisuje vybrané standardní postupy používané k analýze panelových dat uvažující normálně rozdělenou odezvu. Třetí kapitola pokračuje přehledem „novějších" (vesměs však již minimálně 20 let používaných) metod analýzy panelových dat umožňujících uvažovat též nenormálně rozdělenou odezvu. Čtvrtá kapitola uvádí částečný výčet softwarových možností systému R v analýze panelových dat. Pátá kapitola následně shrnuje klasické metody shlukování (aplikovatelné na data, kde lze jednotlivá pozorování interpretovat jako body v eukleidovském prostoru pevné dimenze stejné pro všechna pozorování). Dále je zde uveden (zejména na str. 69) relativně vágně podaný výčet shlukovacích metod pro časové řady (z nichž mnohé by bylo možné použít i s panelovými daty). Nakonec (cca od str. 72) je zde uvedena možnost zobecnění vybraných klasických shlukovacích metod na situaci panelových dat, která však (pokud jsem text správně pochopil) musí být pravidelná. To jest, je vyžadován shodný počet pozorování na panel a též pozorování prováděná pro všechny panely ve stejných časových okamžicích. Panelová/longitudinální pozorování u jednoho subjektu lze potom stále chápat jako bod v eukleidovském prostoru pevné dimenze. Nebo jsem nepochopil text správně a panelová data mohou být i nepravidelná? Lze říci, že kapitoly 1–5 jsou přípravnými a pokoušejí se zejména shrnout známé poznatky, ze kterých se dále vychází. Hlavní teoretický přínos práce lze potom nalézt v šesté a v první polovině sedmé kapitoly kde je navržena nová možnost shlukování na základě panelových dat. Druhá část sedmé kapitoly je věnována popisu simulačních studií, na nichž jsou následně studovány vlastnosti navržených postupů. V poslední osmé kapitole je navržen způsob, jak pomocí shlukování tak zvaně zpřesnit odhady parametrů regresních modelů pro panelová data.

K práci mám tyto zásadnější koncepční výhrady.

1. Rešerše dostupných jak metodologických přístupů, tak souvisejících softwarových implementací použitelných pro shlukování panelových, resp. longitudinálních dat, je značně neúplná. Např. na str. 15 se píše *„Nabídka softwarových nástrojů pro shlukování mikropanelových dat je omezena. K dispozici je pouze jedna funkce, která je implementována v jazyce R v balíčku* `kml`*, autorem je Genolini a Falissard (2011).*" Toto je silně zavádějící tvrzení! Jenom v jazyce R existuje celá řada balíčků rozličných autorů, pomocí nichž lze provádět shlukování na základě panelových dat, viz např. rešerše uvedená v úvodu práce Komárek a Komárková (2014). Na tomto místě chci zdůraznit, že mi nevadí opomenutí práce mé a spoluautorky (nikdy nelze uvést vše). Vadí mi úplné ignorování prací celé oblasti a tvrzení, že „*k dispozici je pouze JEDNA funkce*". Odkaz na svůj článek zde uvádím pouze jako možnost, kde lze nalézt (jistě neúplný) seznam softwarových implementací relevantních pro téma práce.

   Obdobně, též po metodologické stránce nabývá čtenář z předložené práce dojmu, že téměř nic neexistuje pro shlukování (nepravidelných) panelových dat. V tomto ohledu by jistě bylo možné využít některých metod zmíněných v rešerši uvedené na str. 69. Zde se však mluví o shlukování časových řad a z textu není příliš jasné, jakým způsobem by případně bylo možné použít tu či onu metodu v kontextu (i nepravidelných) panelových dat. Zde bych rád podotknul, že oblast shlukování s panelovými/longitudinálními daty je poměrně aktivní oblastí výzkumu. Pro (opět jistě neúplný) přehled

možných relevantních metod, viz např. úvod práce Komárek a Komárková (2013). Opět zdůrazňuji, že mi nevadí opomenutí mé práce, ale opomenutí celé oblasti výzkumu, která je přitom zcela zásadní pro téma předložené disertační práce.

2. Zavedení smíšeného lineárního modelu v kapitole 2 je poněkud zmatečné a dosti zavádějící. Model s pevnými efekty (FEM na str. 32) je zaveden pro obecný vektor regresorů. Pokračuje se modelem s náhodnými efekty (REM na str. 33), kde je v principu uvažován kromě efektu subjektu pouze jeden další a pouze kategoriální regresor, jehož efekt je opět náhodný. Na začátku oddílu o modelu se smíšenými efekty (MEM na str. 34, vzorec 2.11) se přidá obecný vektor regresorů, které všechny odpovídají pevným efektům a následně (ve vzorci 2.12 na str. 35) se tvrdí, že vzorec (2.11) se maticově napíše v dnes standardním tvaru ve značení Lairdové a Warea. Chci podotknout, že vzorec (2.12) pokrývá mnohem širší rodinu modelů než vzorec (2.11), což z autorova textu není příliš zřejmé. Celkově celá část shrnující teorii lineárních smíšených modelů je plná nepřesností až chyb, viz též položky 1–8 seznamu méně zásadních připomínek níže.

3. Kapitola 3 by v práci raději vůbc neměla být. Pro hlavní přínos práce (kapitoly 6 a 7) téměř vše zde uvedené není potřeba (pominu-li několik zmínek o GLMM v kapitole 7, které však není nikde aplikováno). S ohledem na značnou obsáhlost oblasti, kterou se autor v kapitole 3 snaží shrnout, není možné tuto smysluplně popsat na osmi stranách. V důsledku toho je text kapitoly 3 značně povrchní, opět dosti zavádějící a s chybami, viz např. položky 9–13 seznamu méně zásadních připomínek níže.

4. Popis autorem navržených metod (v kapitolách 6 a 7) je takový, že mi není úplně jasné, zda lze vše použít i pro nepravidelná panelová data. Vše podstatné je patrně skryto v úvodu kapitoly 7, kde autor mluví o dvoukrokovém přístupu, kde v prvním kroku se počítají hodnoty navržených shlukovacích proměnných (patrně se jedná o zvolené charakteristiky patřičného smíšeného modelu) a následně se použije některá z konvenčních shlukovacích metod. Toto je poměrně klíčová část pro celou práci a je škoda, že je pojata takto stručně a krypticky.

5. Metodu „zpřesnění odhadů regresních modelů shlukovou analýzou“ bych si nedovolil nazývat jako „zpřesnění odhadů“. Jedná se o to, že po zahrnutí indikátoru příslušnosti do shluku je odhadován *jiný* model. Lze možná mluvit o získání modelu, který lépe vystihuje data, nicméně parametry tohoto modelu mají jinou interpretaci než parametry původního modelu a tedy nejde o žádné zpřesňování odhadů, neboť odhadujeme jiné veličiny.

Práce obsahuje jisté množství vyjádření, o kterých si myslím, že jsou minimálně zavádějící, občas téměř chybná. Jedná se např. o následující.

1. Str. 14: *S regresními modely určenými pro mikropanelová data je spjato úskalí v nejednoznačnosti odhadu jejich parametrů. Regresní parametry a jejich variabilita jsou odhadovány iterativně numerickými optimalizačními metodami. Odhady parametrů modelu a odhady hodnot vysvětlované proměnné jsou nepřesné, pokud kovarianční struktura náhodných chyb (teoretických reziduí) není známá, nebo se v datech vyskytuje směs rozdělení.*
   Hluboce nesouhlasím s tvrzením, že odhady (jakéhokoliv modelu) jsou nejednoznačné, jestliže jsou tyto počítány numericky a iteračně. Samozřejmě, různé iterační algoritmy mají různou přesnost, v důsledku čehož různé algoritmy, resp. různý software produkují numericky odlišné odhady. Toto však nemá co do činění s jednoznačností odhadu (= nějak definovaná statistika neboli funkce dat). Ten buď je anebo není jednoznačný bez ohledu na to, jakou numerickou metodou je počítán. Taktéž nesouhlasím s tvrzením, že odhady parametrů jsou nepřesné, pokud něco. **Všechny** odhady jsou nepřesné a tuto nepřesnost ve statistice vyjadřujeme pomocí směrodatné chyby odhadu.

2. Na str. 29 se hovoří o dvoukrokové analýze, u které se uvádí, že je stále využívána s odkazem na články Johnston a kol. (2008) nebo De Stafano a kol. (2015). Oba tyto články jsou zcela aplikační publikované v medicínských časopisech. Důvodně se domnívám, že ani na jednom z těchto článků se nepodílel autor se statistickým vzděláním. Dvoukrokové odhadování smíšených modelů by skutečný statistik již 20–25 let (cca od publikace článku Lairdové a Warea v roce 1982) nepoužil ani omylem.

3. Náhodná chyba $\varepsilon_{i,t}$ na str. 35 se neřídí *vícerozměrným*, ale obyčejným (*jednorozměrným*) normálním rozdělením.

4. Na str. 36 autor tvrdí, že MEM si poradí s chybějícími hodnotami měření a odvolává se na práci Little a Rubin (2002). Technicky si MEM skutečně poradí s chybějícími hodnotami (tj. vždy dostaneme nějaká čísla), nicméně to, jestli jsou takto získaná čísla (odhady) správná a použitelná, závisí na tom, jakým způsobem tyto chybějící hodnoty vznikají. Právě použitelnosti výsledků se kromě jiného věnuje Little a Rubin.

5. Na konci str. 36 autor píše: „*Pro longitudinální data lze rozlišovat dva základní typy LMM (vzorec 2.11): model s náhodným absolutním členem a model s náhodným sklonem a kombinovat tyto dva typy do jednoho modelu.*“ Zde bych chtěl podotknout, že vzorec (2.11) se pro obecná longitudinální data, kde čas je typicky uvažován spojitě a je modelován nějaký trend v závislosti na čase, příliš nehodí.

6. Za vzorcem (2.15) na str. 37 se píše: „*Model umožňuje modelovat různé tvary individuálních a populačních trajektorií pomocí funkce $g(\lambda_t)$ pro λ_t.*“ Toto nedává smysl. Nemá zde spíše být „pomocí funkce $g(t)$ namísto λ_t“ nebo něco obdobného?

7. V prakticky celé práci se zaměňuje pojem „náhodný efekt“ a „rozptyl“, resp. „komponent rozptylu“ náhodných efektů (např. str. 39 a dále).

8. „*teorií normálního rozdělení*“ na str. 42 se patrně myslí *asymptotická teorie maximální věrohodnosti.*

9. Na konci str. 43 se píše o „*posledních letech*“. Pokud poslední léta znamenají cca 20 let, potom je vše v pořádku. . . .

10. Na začátku str. 44 se píše o doporučení vyzkoušet a porovnat odhady získané z více metod. Zde bych chtěl podotknout, že různé „metody“ (modely) často odpovídají na různé otázky. Mnohdy tedy nemá smysl cokoliv porovnávat. . .

11. Str. 45: Vyjádření „*K odhadu efektů (regresních parametrů) GLMM lze využít metodu ML. Její používání je omezené na malý počet uvažovaných náhodných efektů z důvodu výpočetní náročnosti, kdy se hledá maximum logaritmu věrohodnostní funkci pomocí derivace přes všechny možné hodnoty náhodných efektů.*“ Pomocí „derivace přes všechny možné hodnoty náhodných efektů“ se nic nehledá. . .

12. Stále str. 45: „*Častěji se efekty odhadují pomocí metody REML. Statistici navrhli čtyři metody aproximace věrohodnosti (REML) s cílem umožnit odhadovat složitější GLMM.*“ V kontextu GLMM se žádná metoda REML nepoužívá. Autor má patrně na mysli numerické metody aproximace věrohodnosti.

13. Str. 49: „*MCMC generuje řetězec náhodných výběrů z pravděpodobnostního rozdělení hodnot pevných a náhodných efektů.*“ Potřeba dodat přívlastek „aposteriorního“.

Na závěr textové části posudku bych si dovolil pouze drobnou poznámku k hlavní aplikační oblasti autorova zájmu (výzkum roztroušené sklerózy), zmíněné na konci str. 25. Domnívám se, že zde nejde ani tak o shlukování (klasifikace „bez učitele“), jako spíše o klasifikaci „s učitelem“ (diskriminační analýza), respektive o predikci času do nějaké události (okamžik dosažení disability určitého stupně) na základě longitudinálně

shromažďované informace o nejrůznějších biomarkerech. Předpokládám, že v datech je (u mnoha/všech pacientů) k dispozici informace o stupni klinické disability v jednotlivých okamžicích. V závislosti na tom, jestli je cílem pouze predikovat diabilitu po určité době nebo predikovat (zbývající) čas do dosažení disability jistého stupně, jsou relevantní buď metody longitudinální diskriminační analýzy (pro rešerši možných přístupů viz např. úvod článku Hughes a kol. (2017)), respektive metody tzv. dynamické predikce (viz např. van Houwelingen a Putter (2011) nebo Rizopoulos (2012)).

Závěrem konstatuji, že z práce je vidět, že autor se s nadšením pokusil o aplikační práci z oblasti (zejména lineárních) smíšených modelů, které se pokusil rozšířit směrem využitelným pro ekonomické aplikace. Bohužel však nabývám dojmu, že autorovi chybí řádné základy teorie lineárních smíšených modelů a taktéž širší přehled a povědomí o této oblasti statistiky a taktéž o oblasti klasifikace na základě panelových/longitudinálních dat. V důsledku toho práce obsahuje relativně velké množství neúplností a nepřesností, které mohou vést až k chybnému užití popisovaných modelů nebo k chybné interpretaci výsledků. Navzdory těmto nedostatkům se nicméně domnívám, že zejména v šesté a sedmé kapitole prokázal autor schopnost samostatné vědecké práce, kdy smysluplným způsobem navrhnul inovativní metodu shlukování panelových dat a její vlastnosti následně prozkoumal pomocí obsáhlé simulační studie. Práci proto **lze** dle mého názoru **uznat** za disertační práci na oboru *Kvantitativní metody v ekonomice* a **doporučuji** ji k obhajobě.

Kritéria hodnocení (příloha posudku oponenta)

1. *Odpovídá název práce jeho obsahu (zcela, částečně)?*
   Ano, název práce odpovídá obsahu.

2. *Je vymezení cílů vyhovující, je uvedeno v úvodu, abstraktu?*
   Vymezení cílů je částečně vyhovující. Abstrakt spíše než cíle shrnuje text práce (tj. popisuje, co je v práci obsažno). Jako vyjádření cíle by bylo patrně možné považovat větu: „*V práci je navrženo, jak zpřesnit odhady pomocí shlukové analýzy.*“ O několik řádek níže se píše: „*Dalším cílem práce je zjistit, zda navržený postup shlukování vede ke zkvalitnění regresních modelů pro tento typ dat.*“ Toto již je jasné vymezení cíle. Mám zde pouze mírný problém s chápáním výrazu „*zkvalitnění regresních modelů*“. V jakém smyslu mají být tyto modely kvalitnější? Viz též moje zásadnější připomínka č. 5.

3. *Jak jsou vytčené cíle splněny, je shrnuto v závěru, abstraktu?*
   Cíle jsou částečně splněny návrhem shlukovací metody.

4. *Jsou zřetelně odlišeny metody převzaté z literatury a vlastní přístupy?*
   Ano, jsou.

5. *Co je vědeckým přínosem (např. z hlediska teorie statistiky či metod analýzy dat)?*
   Přínosem je návrh nové shlukovací metody pro panelová data, použitelné v ekonomických aplikacích a vyšetření jejích vlastností pomocí obsáhlé simulační studie.

6. *Úroveň rešerše poznatků a použité literatury ve zkoumané oblasti:*
   Rešerše poznatků z dané oblasti je poměrně problematická, viz též moje zásadnější připomínka č. 1.

7. *Úroveň aplikací, hodnocení a porovnání metod (zařazení tabulek a grafů):*
   Aplikace (zejména v kapitole 7) jsou na přijatelné úrovni. Zejména grafy jsou vhodně voleny a pěkně ilustrují použité metody.

8. *Způsob vyjadřování, jazyková úroveň:*
   Jazyková úroveň práce je na slušné úrovni. Překlepy nebo dokonce gramatické chyby se v práci téměř nevyskytují (mohu-li soudit). Práci bych pouze vytkl relativně vysoký výskyt anglicismů používaných

i v situacích, kdy existuje standardně užívaný český ekvivalent (např. *extenzivně*, *estimátor*, *teorém*). Naopak, pokud je ve větě použito anglické jméno v jiném než prvním pádě, nemělo by evokovat chybné pohlaví nositele (např. na str. 31 „*u Lairda a Wareho*“, kde se mluví o ženě jménem Nan M. Laird).

9. *Jsou řádně definovány používané pojmy, zkratky a symboly?*
   Matematické pojmy a symboly nejsou úplně vždy řádně a jasně definovány (viz též další bod). Zkratky definovány vesměs jsou.

10. *Formální úroveň (matamatických výrazů, tabulek, grafů):*
    Formální úroveň matematických výrazů není příliš dobrá. Mnoho veličin je „definováno“ mezi řečí a vágním způsobem v textu a není úplně vždy jasné, co se vlastně myslí. Tabulky i grafy jsou po formální stránce na výrazně lepší úrovni.

    Dalším formálním nedostatkem, kterého jsem si povšimnul je fakt, že se ne zcela zdařilo setřídění použité literatury dle příjmení autora/autorů (např. reference 68–74 a hned dále 77–78). Se současnými softwarovými možnostmi automatizované tvorby seznamu referencí se jedná o naprosto zbytečný prohřešek.

V Praze dne 4.1.2017

doc. RNDr. Arnošt Komárek, Ph.D.

Reference

Hughes, D., Komárek, A., Czanner, G. a García-Fiãna, M. (2017). Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical Methods in Medical Research (to appear)*, **00**(0), 1–21. doi: 10.1177/0962280216674496.

Komárek, A. a Komárková, L. (2013). Clustering for multivariate continuous and discrete longitudinal data. *The Annals of Applied Statistics*, **7**(1), 177–200. doi: 10.1214/12-AOAS580.

Komárek, A. a Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, **59**(12), 1–38. URL `http://www.jstatsoft.org/v59/i12/`.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press.

van Houwelingen, H. a Putter, H. (2011). *Dynamic prediction in clinical survival analysis*. CRC Press.