Hodnocení dizertační práce v oboru Statistika (VŠE v Praze)

Jméno doktoranda: **Ing. Adam Čabla**

Název práce: **Doba nezaměstnanosti v České republice pohledem analýzy přežití**

Jméno hodnotitele: doc. RNDr. Lenka Komárková, Ph.D.

Posudek oponenta:

Předložená práce je rozdělena do 4 kapitol: Nezaměstnanost, Data, Metodologie a Výsledky. Zmíněným kapitolám ještě předchází Úvod a následuje Závěr. Hlavní text práce bez referencí čítá 107 stran a odkazuje na 120 zdrojů, což je pro tento typ práce adekvátní. Součástí práce je jak textová, tak elektronická příloha v podobě CD. Dizertační práce se zabývá problematikou intervalově cenzorovaných dat, s ní související teorií v podobě analýzy přežití a následnou velmi zajímavou ekonomickou aplikací na datech o nezaměstnanosti.

První kapitola představuje základní statistické ukazatele sledované v oblasti nezaměstnanosti zejména Českým statistickým úřadem (ČSÚ) a Ministerstvem práce a sociálních věcí (MPSV) včetně jejich časového vývoje. Tato kapitola rovněž stručně shrnuje současný stav poznání týkající se případných dopadů a faktorů nezaměstnanosti a následně se pak věnuje rozboru dosavadního výzkumu samotné doby nezaměstnanosti, což je klíčový ukazatel v celé práci.

Druhá kapitola se věnuje datům z Výběrového šetření pracovních sil (VŠPS) získaných od ČSÚ. Představuje detailně jednotlivé proměnné včetně jejich kódování. Zavádí 6 výběrových souborů vzhledem k časovému období – před krizí, během krize a po krizi a povahy respondentů – nezaměstnaní, úspěšní v hledání práce a dále ještě 3 specifické vzorky týkající se nezaměstnaných v posledním čtvrtletí let 2008, 2010 a 2014.

Třetí kapitola prezentuje statistické metody používané v analýze přežití. Nejprve představuje funkci přežití a (kumulativní) rizikovou funkci, typy cenzorování (zleva, zprava, intervalové) s důrazem na intervalově cenzorovaná data. Následně se věnuje odhadům funkce přežití, vícevýběrovému testování shody rozdělení dob do události a používaným regresním modelům v kontextu analýzy přežití (PH model proporčních rizik a AFT model zrychleného času). V části 3.8 pak doktorand seznamuje čtenáře, jak přistupuje k sestavení modelu s větším počtem vysvětlujících proměnných (prediktorů). Závěr této metodologické kapitoly je věnován teorii extrémních hodnot.

Stěžejní část práce lze spatřit v kapitole 4, kdy představená teorie je aplikována na reálných datech z VŠPS. Nejprve je zde odhadnuto celkové pravděpodobnostní rozdělení doby nezaměstnanosti, resp. doby do nalezení práce v jednotlivých časových obdobích včetně jejich následného srovnání vzhledem k vybraným časovým obdobím. Konkrétně byl použit Turnbullův neparametrický odhad v kombinaci s logrankovým testem, parametrický AFT model a parametrický odhad vycházející z logaritmicko-logistického rozdělení. Část zabývající se modelováním funkce přežití je zakončena inovativním přístupem domodelování jejího pravého konce od předem stanovené prahové hodnoty, tj. funkce přežití je odhadnuta po částech. Autor na datech dokládá vyšší přesnost takto provedených odhadů.

Další část této kapitoly 4.3 se zabývá párovými vztahy mezi sledovaným ukazatelem doby nezaměstnanosti, resp. doby do nalezení práce a 16 vybranými proměnnými, jako je např. pohlaví, věk, vzdělání respondenta, jeho registrace na Úřadu práce, či způsob a počet metod použitých při hledání práce. Navazující části 4.4 a 4.5 se věnují vybudování vhodného co nejjednoduššího vícenásobného AFT modelu, tzv. minimálního adekvátního modelu, vyjadřujícího co nejlépe vztah mezi sledovanou odezvou a vhodnou podmnožinou prediktorů včetně redukce kategorií, nejprve bez uvažovaných vzájemných interakcí a následně pak i s nimi. K zmíněné analýze byl využit statistický software R (https://www.r-project.org/).

Na práci kladně hodnotím, že doktorand řešil danou problematiku komplexním způsobem. Na nezaměstnanost se zaměřil ve třech ekonomicky odlišných obdobích. Vytvořil několik vzorků z důvodu zvýšení vypovídající schopnosti výsledků. Pro analýzu použil jak parametrické, tak neparametrické statistické metody. Zkombinoval poznatky z dvou oblastí, a to analýzy přežití a teorie extrémních hodnot. Je na místě rovněž zdůraznit, že v české odborné literatuře není intervalově cenzorovaným datům věnován takový prostor, a to ani v přírodních vědách, jako v případě dat cenzorovaných zprava.

Ač je práce velmi zajímavá, není dle mého názoru psána příliš pečlivě. Práce obsahuje zbytečné překlepy, jako např. *„rozdělní“* místo *„rozdělení“*, viz str. 7 a popisky tabulek 4.5, 4.8, 4.11, jazykové prohřešky s velkým a malým písmenem jako *„studentovo rozdělení“* (str. 46) či *„Chí-kvadrát rozdělení“* (str. 36), špatné skloňování (*„pro souboru“*, viz str. 96), nejednotnost termínů *„model s urychleným selháním“* (str. 33) vs. *„model zrychleného času“* a symbolů (〉 vs. ( ] pro polouzavřený interval, resp. *log* a *ln* pro přirozený logaritmus. Občas zkratka (např. VŠPS) předchází jejímu zavedení u příslušného názvu/pojmu, viz str. 8 a 9. Při odkazování na původní zdroje není brán zřetel na český jazyk, např. *„navržený v Nelson (1972)“*, *„ve Vojtěch (2010)“*. V seznamu literatury u nemalé části časopiseckých publikací chybí stránky, např. hned na první straně seznamu je tento nedostatek u 6 zdrojů (1, 4, 6, 7, 10, 13), a navíc zde není ani jednotný styl. U publikace č. 60 není pro změnu uveden název konferenčního příspěvku.

V práci lze nalézt i překlepy (resp. chyby) v matematických výrazech, viz např. symbolický zápis alternativní hypotézy trendu (3.31) a (3.32). Dojem z práce rovněž kazí některé formulace „*odhadnout NPMLE“*, tj. odhadnout odhad, nevysvětlené symboly např. u vzorce (3.11), chybějící symboly, jako je tomu ve vzorcích (3.45) a (3.50), kde ve jmenovateli u druhého z vektorů chybí symbol transpozice, či v případě matematického vzorce (3.66) na str. 43, kde u lineárního regresního modelu postrádám buďto náhodnou složku ε, nebo symbol střední hodnoty E před odezvou Y. Navíc celé následné vysvětlení tohoto modelu s interakcí (v rámci textu ve formě odrážek) je nepřesné až spíše chybné. Zavádějící je rovněž tvrzení na str. 35, že *„Nevýhodou plně parametrického modelu je omezení na dvě pravděpodobnostní rozdělení – exponenciální a Weilbullovo“* – tato dvě rozdělení jsou možná a hojně užívaná, nikoliv však jediná možná.

Co se týče prezentace výsledků, uvítala bych delší popisky u tabulek a grafů, neboť ty by měly být čitelné čtenáři bez delšího studování textu. Ocenila bych rovněž vyšší formální úroveň předložených grafů, zejména v části 4.3, kde se používá typově stejný graf. Proč není u daných grafů v rámci stejného obrázku použito stejné měřítko na horizontální ose? Bylo by tak možné lépe srovnávat. A proč toto měřítko není voleno vhodněji (s použitím exponenciální transformace) kvůli následným interpretacím? Dále bych uvítala na vertikální ose alespoň drobnou mezeru mezi popisky, které odpovídají různým časovým obdobím. Nevím, zda je šťastné do grafu zahrnovat odhad pro referenční skupinu, i když chápu, že referenční skupina někde být uvedena musí a toto je možný způsob.

Vítám, že autor práce pouze nekonstatuje holé výsledky statistických analýz, ale snaží se i o další interpretaci či případný komentář související s „nestatistickou" problematikou. Na druhou stranu však některé závěry jsou poněkud nešťastně formulovány (str. 82: *„Svou roli může hrát vliv vzdělání, kdy mladší lidé jsou obecně vzdělanější, ale neplatí to pro nejmladší kategorie, zejména pro kategorie 15 – 19 let a 20 – 24 let."*), či obsahují dle mého názoru nejasnou dedukci (str. 82: *„Změny v době nezaměstnanosti v průběhu času jsou vysvětlitelné vlivem propouštění – např. zkrácení doby nezaměstnanosti v době krize, zejména u starších osob, u kterých může být propouštění masivnější."*).

Vícenásobným modelům (části 4.4 a 4.5) není věnován dostatečně velký prostor. V textu jsem nenašla zmínku, které prediktory se ukázaly jako statisticky nevýznamné, přestože z obrázku 4.35 vyplývá, že některé z nich se ve výsledném modelu neobjevily. Rovněž by nebylo na škodu v těchto dvou kapitolách použít nějakou tabulku pro výsledný model. Tento nedostatek částečně vyvažuje představený algoritmus tvorby minimálního adekvátního modelu v části 3.8, ze kterého je zřejmé, že doktorand nepoužil při postupném vyřazování prediktorů z plného modelu jediného kritéria, ale vzal do úvahy i další aspekty včetně kombinace zpětného a dopředného krokování.

Závěrem konstatuji, že práce prezentuje řadu statistických analýz vyžadující znalosti z pokročilejších partií statistických metod, navíc má potenciál pro případný budoucí výzkum. Dále je třeba ocenit i aktuálnost tématu z hlediska hospodářské politiky. Doktorand danou prací prokázal, že je schopen samostatné vědecké práce, což dokládá jím navržený přístup modelování hlavní části a pravého konce rozdělení doby do události. Předložená práce dle mého názoru splňuje nároky kladené na dizertační práce na Vysoké škole ekonomické v Praze, a proto ji **doporučuji k obhajobě** před příslušnou komisí. Dále k tomuto posudku přikládám návrh na dotazy k obhajobě a vyplněný dotazník – kritéria hodnocení.

V Jindřichově Hradci dne 30. dubna 2017

doc. RNDr. Lenka Komárková, Ph.D.

Příloha posudku oponenta č. 1:

Dotazy k obhajobě:

1. Jaký je rozdíl mezi Vámi uvedeným náhodným a nezávislým cenzorováním? Podrobněji vysvětlete.
2. Mohl byste uvést možné příklady informativního cenzorování a může tento typ cenzorování teoreticky nastat v kontextu Vašich dat?
3. V případě kvantitativního prediktoru *„vek“* jste se rozhodl na základě porovnání věrohodností obou zvažovaných modelů pro jeho kategoriální variantu *„veksk“*. Čím si tento jev vysvětlujete a šlo by postupovat i jinak se zachováním kvantitativní povahy, tj. plné informace, této proměnné?
4. Jak byste u modelů s jedním kvalitativním prediktorem, se kterými v části 4.3 zejména pracujete, provedl dodatečné mnohonásobné porovnávání mezi jednotlivými kategoriemi tohoto prediktoru?
5. V případě delší longitudinální studie, by se Vám mohlo stát, že některé statistické jednotky budou nezaměstnané opakovaně, jak byste k tomuto problému přistoupil?

Příloha posudku oponenta č. 2:

Kritéria hodnocení:

1. ***Odpovídá název práce jeho obsahu (zcela, částečně):***
   Ano, název dizertační práce zcela odpovídá jejímu obsahu.
2. ***Je vymezení cílů vyhovující, je uvedeno v úvodu, abstraktu?***
   Ano, vymezení cílů je vyhovující.
   Cíle jsou představeny jak v abstraktu, tak podrobněji v úvodu práce.
3. ***Jak jsou vytčené cíle splněny, je shrnuto v závěru, abstraktu?***
   Vytčené cíle jsou splněny. V závěru jsou pěkně shrnuty. Abstrakt se spíše dle mého názoru věnuje zdůraznění přínosů práce, než shrnutí hlavních výsledků.
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 dizertační práce je sestavení vícenásobného minimálního adekvátního AFT modelu pro dobu nezaměstnanosti a vlastní přístup k modelování pravděpodobnostního rozdělení doby do nalezení práce, kdy odhad pravého konce tohoto rozdělení vychází z teorie extrémních hodnot.
6. ***Úroveň rešerše poznatků a použité literatury ve zkoumané oblasti:***
   Dostatečná. Práce se odkazuje na 120 zdrojů. Nicméně, v seznamu literatury u nemalé části časopiseckých publikací chybí stránky.
7. ***Úroveň aplikací, hodnocení a porovnání metod (zařazení tabulek a grafů):***
   Aplikace jsou velmi pěkné a zajímavé. Tabulky a grafy doplňují vhodně daný text.
8. ***Způsob vyjadřování, jazyková úroveň:***
   Jazyková úroveň není úplně dobrá. V práci se objevují neobratné formulace, občas i hovorové výrazy. Nešťastně je z jazykového hlediska použit styl odkazování na zdroje.
9. ***Jsou řádně definovány používané pojmy, zkratky a symboly?***
   Většinou ano. Na konci práce je uveden samostatný seznam zejména matematických symbolů.
10. ***Formální úroveň (matematických výrazů, tabulek, grafů):***
    Formální úroveň výrazů, tabulek a grafů je na přijatelné úrovni. Některé matematické výrazy nejsou formulovány zcela přesně. Dále pak prezentované grafy by mohly být názornější.

V Jindřichově Hradci dne 30. dubna 2017

doc. RNDr. Lenka Komárková, Ph.D.