VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

Doba nezaměstnanosti v České republice pohledem analýzy přežití

Disertační práce

Doktorand: Ing. Adam Čabla

Školitelka: doc. RNDr. Ivana Malá, CSc.

Studijní program: Kvantitativní metody v ekonomice

Studijní obor: Statistika

Praha, březen 2017

Prohlášení

Prohlašuji, že jsem disertační práci zpracoval samostatně a že jsem uvedl všechny použité prameny a literaturu, ze kterých jsem čerpal.

V Praze dne 15. 3. 2017

…….

podpis

Poděkování

Rád bych na tomto místě poděkoval své školitelce, doc. RNDr. Ivaně Malé, CSc. za dlouholetou trpělivost, pomoc a cenné rady.

Dále pak své ženě za trpělivost.

Děkuji.

Abstrakt

Předkládaná disertační práce má za cíl aplikovat metody analýzy přežití na data z Výběrového šetření pracovních sil, která jsou intervalově rozdělená. S ohledem na to používám v práci specifické metody určené pro práci s tímto typem dat, zejména Turnbullův odhad, vážený log-rank test a AFT model.

Dalšími cíli práce jsou návrh a aplikace metodiky pro tvorbu modelu doby do nalezení práce v závislosti na dostupných faktorech a jeho interpretaci, popis vývoje pravděpodobnostní rozdělení doby nezaměstnanosti a zpřesnění odhadu pravého konce pomocí aplikace teorie extrémních hodnot.

Mezi hlavní přínosy práce můžu zařadit tvorbu metodiky pro zkoumání dat z Výběrového šetření pracovních sil založenou na standardních postupech analýzy přežití. Jelikož jsou tato data mezinárodně srovnatelná, je tato metodika použitelná na úrovni zemí Evropské Unie a některých dalších. Dalším přínosem práce je odhad parametrů zobecněného Paretova rozdělení na intervalově cenzorovaných datech a tvorba a srovnání různých modelů po částech spojené distribuční funkce včetně řešení problému navázání dvou distribučních funkcí.

Práce přinesla empirické výsledky, z nichž nejzásadnější je srovnání výsledků ze tří různých datových přístupů a konkrétní vztahy mezi vybranými faktory a dobou do nalezení práce či dobou nezaměstnanosti.

Klíčová slova: Analýza přežití, Výběrové šetření pracovních sil, intervalové cenzorování, teorie extrémní hodnoty, doba nezaměstnanosti

Abstract

In the presented thesis the aim is to apply methods of survival analysis to the data from the Labour Force Survey, which are interval-censored. With regard to this type of data, I use specific methods designed to handle them, especially Turnbull estimate, weighted log-rank test and the AFT model.

Other objective of the work is the design and application of a methodology for creating a model of unemployment duration, depending on the available factors and its interpretation. Other aim is to evaluate evolution of the probability distribution of unemployment duration and last but not least aim is to create more accurate estimate of the tail using extreme value theory.

The main benefits of the thesis can include the creation of a methodology for examining the data from the Labour Force Survey based on standard techniques of survival analysis. Since the data are internationally comparable, the methodology is applicable at the level of European Union countries and several others. Another benefit of this work is estimation of the parameters of the generalized Pareto distribution on interval-censored data and creation and comparison of the models of piecewise connected distribution functions with solution of the connection problem.

Work brought empirical results, most important of which is the comparison of results from three different data approaches and specific relationship between selected factors and time to find a job or spell of unemployment.

Keywords: Survival Analysis, Labour Force Survey, Interval Censoring, Extreme Value Theory, Unemployment Duration

Obsah

Abstrakt 3

Abstract 4

Obsah 5

Úvod 7

1 Nezaměstnanost 9

1.1 Definice 9

1.2 Statistiky nezaměstnanosti 10

1.3 Výběrové šetření pracovních sil 13

1.4 Dopady a faktory nezaměstnanosti 16

2 Data 19

2.1 Doba nezaměstnanosti 20

2.2 Vysvětlující proměnné 21

3 Metodologie 25

3.1 Analýza přežití 25

3.2 Cenzorování 27

3.2.1 Intervalové cenzorování I. typu 29

3.2.2 Intervalové cenzorování II. typu 29

3.2.3 Dvojité cenzorování 29

3.2.4 Panel count data 30

3.3 Turnbullův odhad 30

3.4 Vážený log-rank test 32

3.5 Model zrychleného času 35

3.6 Úsudky v rámci modelu zrychleného času 36

3.7 Log-logistické rozdělení 38

3.8 Budování regresního modelu 40

3.9 Teorie extrémní hodnoty 43

3.9.1 Rozdělení výběrového maxima 44

3.9.2 Rozdělení špiček nad prahem 45
3.9.3 Po částech spojená distribuční funkce 47
3.10 Software 45
4 Výsledky 50
4.1 Celkové rozdělení 53
4.2 Rozdělení konce 63
4.3 Jednotlivé proměnné 63
4.3.1 Regiony 71
4.3.2 Počet osob v domácnosti 74
4.3.3 Vztah k osobě v čele domácnosti 75
4.3.4 Pohlaví 77
4.3.5 Rodinný stav 78
4.3.6 Věk 78
4.3.7 Země narození 81
4.3.8 Státní příslušnost 83
4.3.9 Zdravotní postižení 84
4.3.10 Vzdělání 84
4.3.11 Registrace na Úřadu práce 86
4.3.12 První zaměstnání 87
4.3.13 Typ hledaného úvazku 89
4.3.14 Metody hledání 90
4.3.15 Počet užitých metod hledání zaměstnání 93
4.3.16 Velikost obce 95
4.4 Minimální adekvátní model 96
Závěr 101
Seznam literatury a zdrojů 108

Úvod

Nezaměstnanost je jedním z nejvíce diskutovaných témat v oblasti ekonomické teorie a politiky. Ze statistického pohledu se problém nezaměstnanosti obvykle popisuje charakteristikami rozsahu nezaměstnanosti, jako například obecná a specifické míry nezaměstnanosti či podíl nezaměstnaných osob.

Pokud je vedena diskuze o délce trvání nezaměstnanosti, obvykle se pracuje s pojmem dlouhodobá nezaměstnanost, čímž se míní nezaměstnanost trvající déle než jeden rok. Tato nezaměstnanost se pak opět popisuje výše zmíněnými charakteristikami. Oproti tomu se málo prostoru věnuje délce nezaměstnanosti jako náhodné veličině, u níž můžeme popsat pravděpodobnostní rozdělení s mnohem většími detaily a můžeme pak zkoumat vliv různých faktorů na rozdělení této náhodné veličiny.

Rozdělení doby nezaměstnanosti do intervalů a sledování relativních četností v těchto intervalech provádějí a pravidelně zveřejňují Český statistický úřad a Ministerstvo práce a sociálních věcí. Obě instituce zveřejňují i podrobnější členění, např. podle pohlaví či vzdělání.

Hlavním cílem této práce je podrobně prozkoumat pravděpodobnostní rozdělení doby do nalezení práce a doby nezaměstnanosti. Jedná se o empirickou práci s inovativní aplikací. Dílčími cíli práce jsou:

- aplikace analýzy přežití na data z Výběrového šetření pracovních sil,
- návrh metodiky pro tvorbu modelu doby nezaměstnanosti v závislosti na dostupných faktorech a jeho interpretaci,
- samotný popis vývoje pravděpodobnostního rozdělení doby nezaměstnanosti během krize,
- pokusit se zpřesnit odhad pravého konce pravděpodobnostního rozdělení doby do nalezení práce pomocí aplikace teorie extrémní hodnoty a zhodnotit úspěšnost tohoto zpřesnění,
- v rámci teorie extrémní hodnoty porovnat tři různé přístupy k napojování rozdělení hlavní částí a pravého konce pravděpodobnostního rozdělní.

V práci jsou užita data z Výběrového šetření pracovních sil za tři časově odlišná období, z nichž první zahrnuje čtvrté čtvrtletí roku 2007 a celý rok 2008, druhé celý rok 2010 a první čtvrtletí roku 2011 a třetí obsahuje čtvrté čtvrtletí roku 2013 a celý rok 2014. Toto rozvržení umožňuje zkoumat vývoj doby do nalezení práce a doby nezaměstnanosti ve vztahu k těmto ekonomicky odlišným obdobím.

Údaje z VŠPS jsou intervalově cenzorované, proto jejich zpracování vyžaduje použít specifické metody vycházející z analýzy přežití. Ty zahrnují Turnbullův odhad, neparametrický log-rank test uzpůsobený pro tento typ dat a model zrychleného času. Vedle této standardní metodologie v práci aplikuji model po částech spojené distribuční funkce a směs pravděpodobnostních rozdělení za účelem zjištění chování pravého konce rozdělení doby do nalezení zaměstnání, čímž se dosud nikdo nezabýval.

Práce je rozdělena do čtyř základních částí. V první z nich definuji termín nezaměstnanost a popisuji zdroje a běžně užívané charakteristiky nezaměstnanosti ze strany Českého statistického úřadu, Ministerstva práce a sociálních věcí a OECD. Je zde uveden vývoj vybraných charakteristik a jejich srovnání. V této části také diskutuji základní faktory ovlivňující nezaměstnanost a dopady nezaměstnanosti.

Ve druhé části práce se zabývám podrobným popisem použitých dat, především rozlišením tří datových souborů použitých dále, popisem výpočtu doby nezaměstnanosti z dat Výběrového šetření pracovních sil a popisem faktorů, které se v datech nachází a jejichž vliv na dobu nezaměstnanosti zkoumám.

Ve třetí části popisuji metody užité pro samotné zkoumání. Jsou zde stručně představeny analýza přežití, pojem cenzorovaných pozorování, neparametrické metody (Turnbullův odhad a vážený log-rank test), parametrický model zrychleného času a úsudky v něm založené na metodě maximální věrohodnosti a na závěr uvádím teorii extrémních hodnot a zabývám se modely umožňujícími zpřesnit odhad pravého konce. Součástí této kapitoly je i stručné představení balíčků softwaru R použitých pro jednotlivé výpočty v této práci.

Čtvrtá část obsahuje odhady pravděpodobnostního rozdělení doby do nalezení práce a doby nezaměstnanosti, odhady konce rozdělení doby do nalezení práce, testuji a odhaduji vliv jednotlivých proměnných na dobu nezaměstnanosti, představuji model zahrnující všechny statisticky významné vlivy na dobu nezaměstnanosti najednou v tzv. minimálním adekvátním modelu včetně diskuze vztahu k úplnému modelu a na závěr se zabývám interakcemi vybraných dvojic proměnných ve vztahu k době nezaměstnanosti.

Vzhledem k povaze práce není možné uvádět kompletní výstupy přímo v práci, proto jsou tyto obsaženy v elektronických přílohách. Příloha A obsahuje všechny modely z kapitoly 4.3, příloha B obsahuje informace o některých odhadnutých charakteristikách z těchto modelů, příloha C obsahuje postup výpočtu minimálního adekvátního modelu a modelu s interakcemi v kapitolách 4.4 a 4.5 a příloha D obsahuje četnosti jednotlivých vysvětlujících proměnných.

Na konci práce je též zařazena příloha obsahující obrázky ke kapitole 4.5.

1 Nezaměstnanost

Cílem této kapitoly je shrnout základní informace o nezaměstnanosti, délce nezaměstnanosti, dopadech nezaměstnanosti a zdrojových datech o nezaměstnanosti. Představím i základní statistiky, které se k tématu nezaměstnanosti váží a jejich vývoj v České republice.

1.1 Definice

Nezaměstnanost se dá definovat různými způsoby, zde budu primárně pracovat s nezaměstnanými podle definice Mezinárodní organizace práce (*International Labour Organization* – ILO), viz ILO (1982). Podle této definice je nezaměstnaným každý člověk určitého věku (u nás standardně 15 – 64 let), který

- je bez práce, tedy nebyl v žádném placeném zaměstnání ani sebezaměstnání v referenčním období (u nás 1 týden),
- je připraven nastoupit do práce, tedy připraven nastoupit do placeného zaměstnání nebo sebezaměstnání nejpozději do 14 dnů v referenčním období,
- hledá práci, tedy podnikal určité kroky k nalezení placeného zaměstnání nebo sebezaměstnání v referenčním období (u nás za poslední 4 týdny). Tyto specifické kroky zahrnují například registraci na úřadu práce či v soukromé pracovní agentuře, osobní hledání práce, podávání a odpovídání na inzeráty, hledání pomoci u známých či příbuzných, hledání půdy, budov, strojů nebo vybavení k založení vlastního podniku, hledání finančních zdrojů, žádání o licence apod.

Této definici nezaměstnanosti odpovídají údaje o nezaměstnanosti publikované Českým statistickým úřadem (ČSÚ) na základě Výběrového šetření pracovních sil (VŠPS, viz ČSÚ (2016)), včetně obecné míry nezaměstnanosti, specifických měr nezaměstnanosti a míry dlouhodobé nezaměstnanosti.

Kromě této definice definuje ILO i pojem zaměstnané osoby, což je osoba starší 14 let, která v průběhu referenčního období pracovala alespoň 1 hodinu a za svou práci dostala zaplaceno. Patří sem i sebezaměstnané osoby.

Poslední důležitý pojem je ekonomicky neaktivní osoba, což je dle ILO osoba, která není ani zaměstnaná ani nezaměstnaná. Sem spadají např. děti, studenti, důchodci, ženy v domácnosti a další.

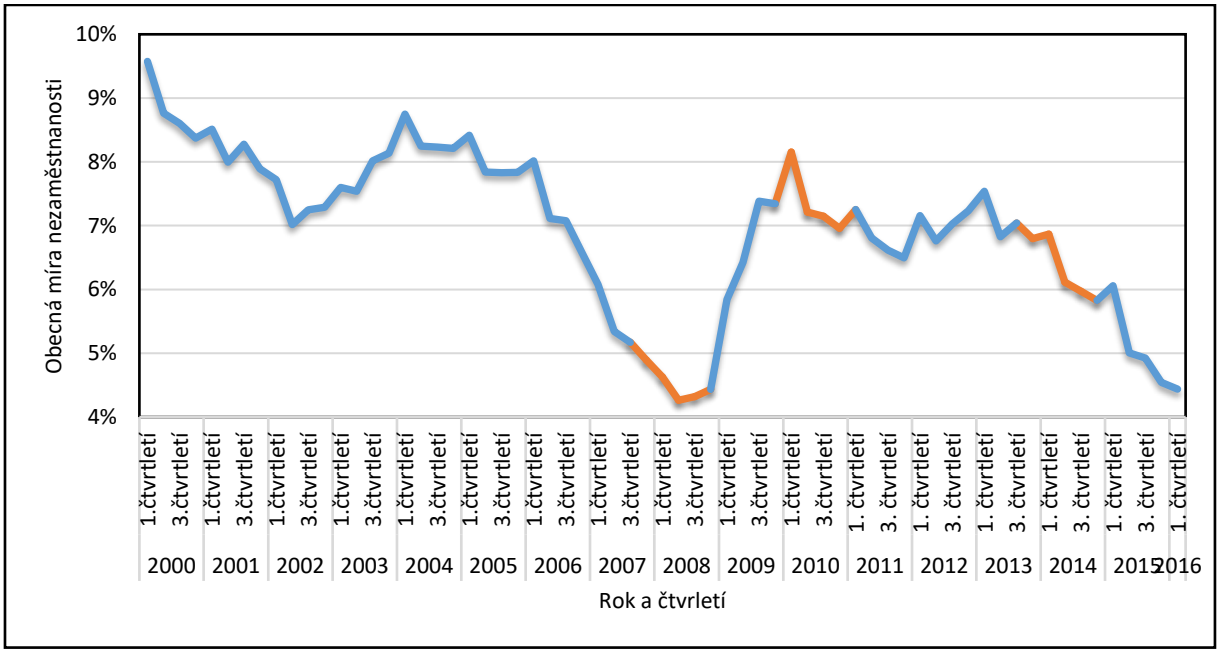
Jiným ukazatelem je podíl nezaměstnaných osob, který je publikován Ministerstvem práce a sociálních věcí (MPSV) na základě údajů z Úřadů práce (ÚP). V tomto případě je za nezaměstnaného považován každý tzv. dosažitelný neumístěný uchazeč o zaměstnání podle zákona o zaměstnanosti č. 435/2004 §24, který může ihned nastoupit do zaměstnání. Jinak řečeno nezaměstnaným je v tomto smyslu každý člověk, který „osobně požádá o zprostředkování vhodného zaměstnání krajskou pobočku Úřadu práce, v jejímž územním obvodu má bydliště, a při splnění zákonem stanovených podmínek je krajskou pobočkou Úřadu práce zařazena do evidence uchazečů o zaměstnání.“[1]

1.2 Statistiky nezaměstnanosti

Níže uvádím veřejně publikované statistiky nezaměstnanosti ze strany ČSÚ a MPSV.

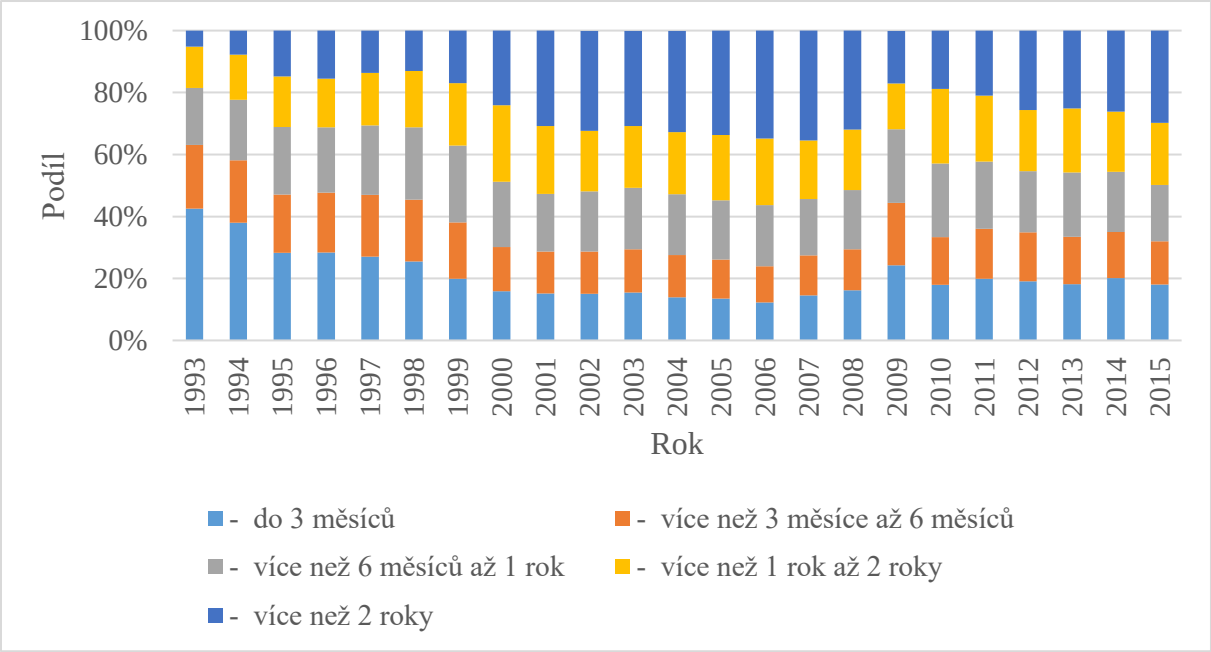
Obecná míra nezaměstnanosti *OBN* je určena jako podíl počtu nezaměstnaných *N* na celkové pracovní síle *EA*. Celková pracovní síla je definovaná jako součet počtu nezaměstnaných a počtu zaměstnaných *Zam*, tedy v terminologii ILO je to počet ekonomicky aktivních osob. Můžeme psát

$$OBN = \frac{N}{EA} = \frac{N}{N + Zam}. \tag{1.1}$$



Obrázek 1.1 Čtvrtletní obecná míra nezaměstnanosti v České republice, zdroj: ČSÚ

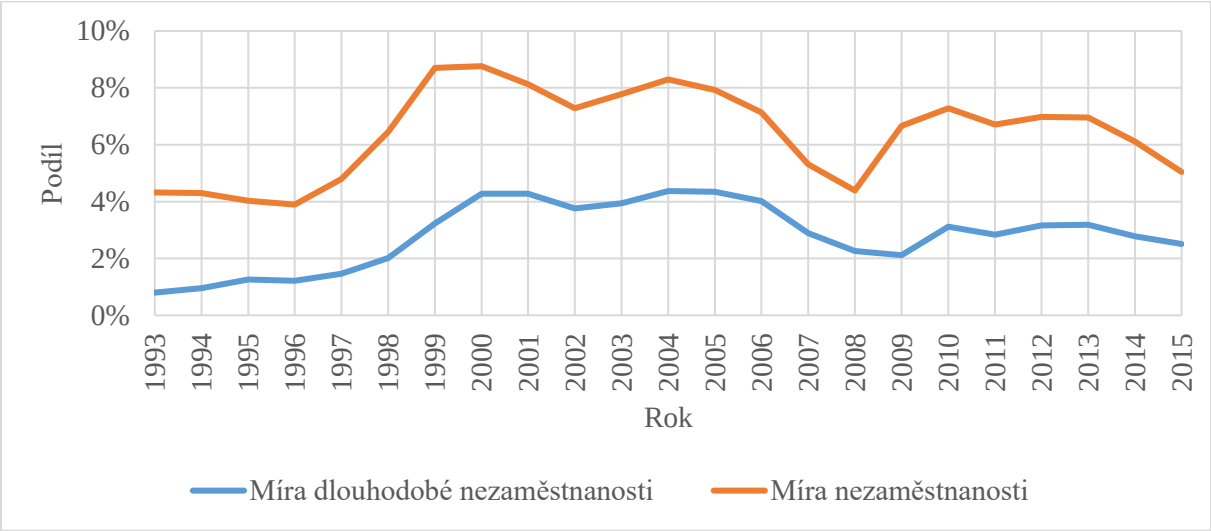
[1] https://portal.mpsv.cz/sz/obecne/prav_predpisy/akt_zneni/ZOZ_PLATNE_ZNENI_OD_1.9.2016.PDF



Obrázek 1.2 Roční rozdělení doby nezaměstnanosti v České republice, zdroj: ČSÚ

Na obrázku 1.1 je vidět vývoj obecné míry nezaměstnanosti v České republice od roku 2000 do poloviny roku 2016. Červeně jsou vyznačená období použitá v této práci (více v kapitole 2).

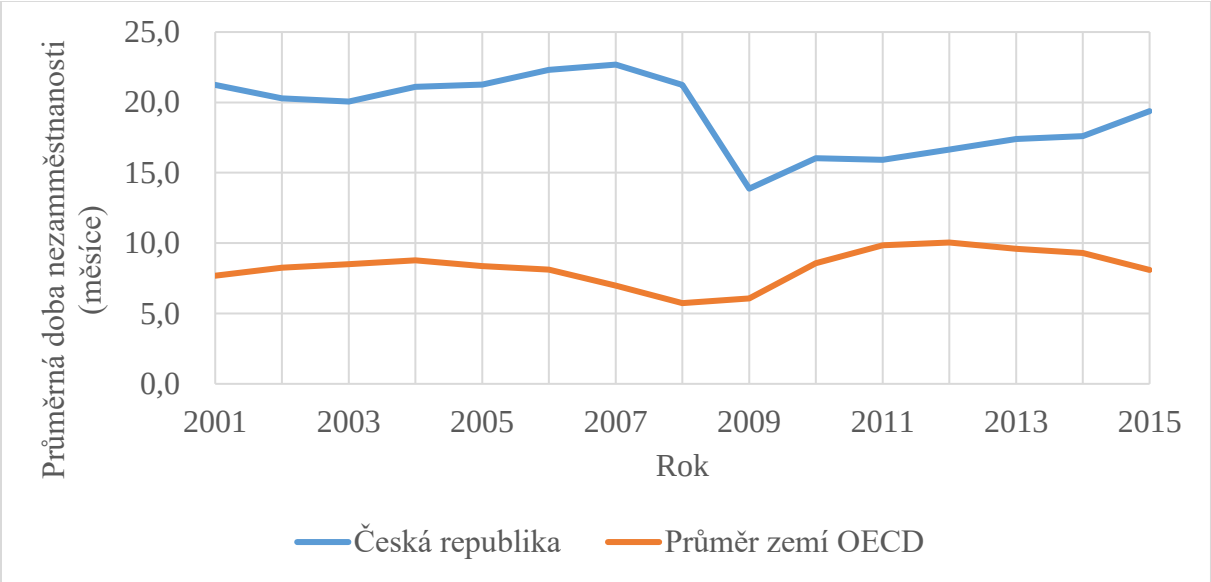
Specifické míry nezaměstnanosti se počítají stejně jako obecná míra s tím, že se konstruují pro jasně definované specifické skupiny obyvatel, např. pro muže a ženy, pro různé stupně vzdělání, pro různé věkové kategorie apod.



Obrázek 1.3 Roční míry nezaměstnanosti a dlouhodobé nezaměstnanosti v České republice, zdroj: ČSÚ

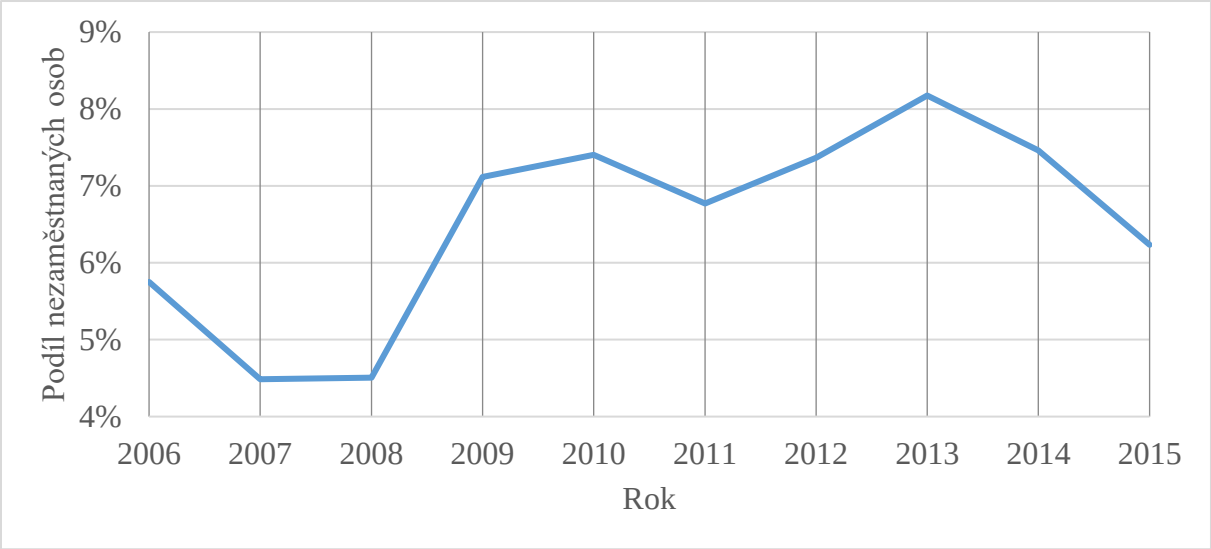
Podíl dlouhodobě nezaměstnaných se počítá jako podíl osob nezaměstnaných déle než jeden rok na celkovém počtu nezaměstnaných. Trochu odlišným ukazatelem je pak míra dlouhodobé nezaměstnanosti, která vyjadřuje podíl počtu nezaměstnaných déle než jeden rok na celkové pracovní síle. Vývoj těchto ukazatelů v letech 1993 – 2015 v České republice zachycují obrázky 1.2 a 1.3.

Průměrná doba nezaměstnanosti je ukazatel sledovaný OECD (2016a) opět na základě dat VŠPS (OECD, 2016b). Obrázek 1.4 zachycuje vývoj průměrné doby nezaměstnanosti v letech 2001 až 2015 pro Českou republiku a průměr členských zemí OECD.



Obrázek 1.4 Průměrná doba nezaměstnanosti v České republice a OECD, zdroj: OECD

Podíl nezaměstnaných osob je ukazatel zveřejňovaný MPSV, který se počítá jako podíl počtu nezaměstnaných registrovaných na ÚP a celkového počtu obyvatel ve věku 15 – 64 let včetně. Tento ukazatel se užívá od začátku roku 2013, předchozím ukazatelem MPSV byla registrovaná míra nezaměstnanosti, který se určovala jako podíl počtu nezaměstnaných registrovaných na ÚP a ekonomicky aktivních osob. Je zřejmé, že počet ekonomicky aktivních osob je nižší než počet všech osob ve věku 15 – 64 let, tedy dříve užívaná registrovaná míra nezaměstnanosti nabývala vyšších hodnot než současný ukazatel. Roční podíl nezaměstnaných osob v České republice dopočítaný za období 2006 až 2015 je zobrazen na obrázku 1.5. Více informací je např. ve společné tiskové zprávě ČSÚ a MPSV (2012).



Obrázek 1.5 Podíl nezaměstnaných osob v České republice, zdroj: MPSV

1.3 Výběrové šetření pracovních sil

VŠPS je šetřením, které provádí ČSÚ od prosince 1992 kontinuálně v průběhu celého roku. Jeho hlavním cílem je „získávání pravidelných informací o situaci na trhu práce, umožňujících její analýzu z různých hledisek, zejména ekonomických, sociálních a demografických." (ČSÚ, 2015). Od roku 2002 je dotazník VŠPS plně harmonizován se standardy Eurostatu a jedná se o národní modifikaci *Labour Force Survey* (LFS). LFS je velké výběrové šetření domácností, které přináší čtvrtletní výsledky o podílu obyvatel na pracovní síle. LFS se provádí ve všech zemích Evropské Unie a Evropské asociace volného obchodu (*Eurpean Free Trade Association* – EFTA) a přináší tak statistiky srovnatelné napříč zeměmi ve shodě s metodikou ILO, více na Eurostat (2016).

Při VŠPS se používá dvoustupňový náhodný výběr, kdy jednotkou prvního výběru je sčítací obvod, který se vybírá metodou znáhodněného systematického výběru s pravděpodobnostmi zahrnutí odpovídajícími počtu trvale obydlených bytů. Jednotkou druhého výběru je byt, který je ze sčítacího obvodu vybrán metodou prostého náhodného výběru. Oporou výběru je registr sčítacích obvodů a budov.[2]

Předmětem šetření jsou všechny osoby obvykle bydlící v domácnostech šetřených bytů, které setrvávají či mají v úmyslu zůstat na území České republiky alespoň jeden rok. U všech osob starších 14 let se zjišťují podrobné údaje o ekonomické aktivitě a vzdělání.

[2] http://apl.czso.cz/irso4/

VŠPS se koná kontinuálně v průběhu celého roku, přičemž údaje se typicky shrnují po jednotlivých čtvrtletích a každý byt je šetřen právě jednou za čtvrtletí. Každý byt setrvává v souboru po dobu 5 po sobě jdoucích šetření, soubor se tedy obměňuje každé čtvrtletí z 20 %.

Jednotlivým osobám se přiřazují váhy určené jako podíl počtu osob v celé populaci a počtu osob ve výběru stejného pohlaví ve stejné věkové skupině a okresu bydliště s výjimkou Prahy, která se bere jako jeden celek. Počty osob v celé populaci jsou projektovány z koncového stavu předchozího roku na střed aktuálního čtvrtletí pomocí přirozeného přírůstku a úbytku a salda migrace v předcházejícím roce. (ČSÚ, 2013)

Využívání dat z VŠPS má oproti datům z MPSV následující výhody:

- Údaje jsou mezinárodně srovnatelné s údaji jiných evropských zemí.
- Lze jej považovat za náhodný výběr nezaměstnaných dle mezinárodně uznávané definice ILO.
- Jsou sledovány údaje za celé byty respektive domácnosti v bytech, což umožňuje sledovat více proměnných, než se nachází v datech MPSV.

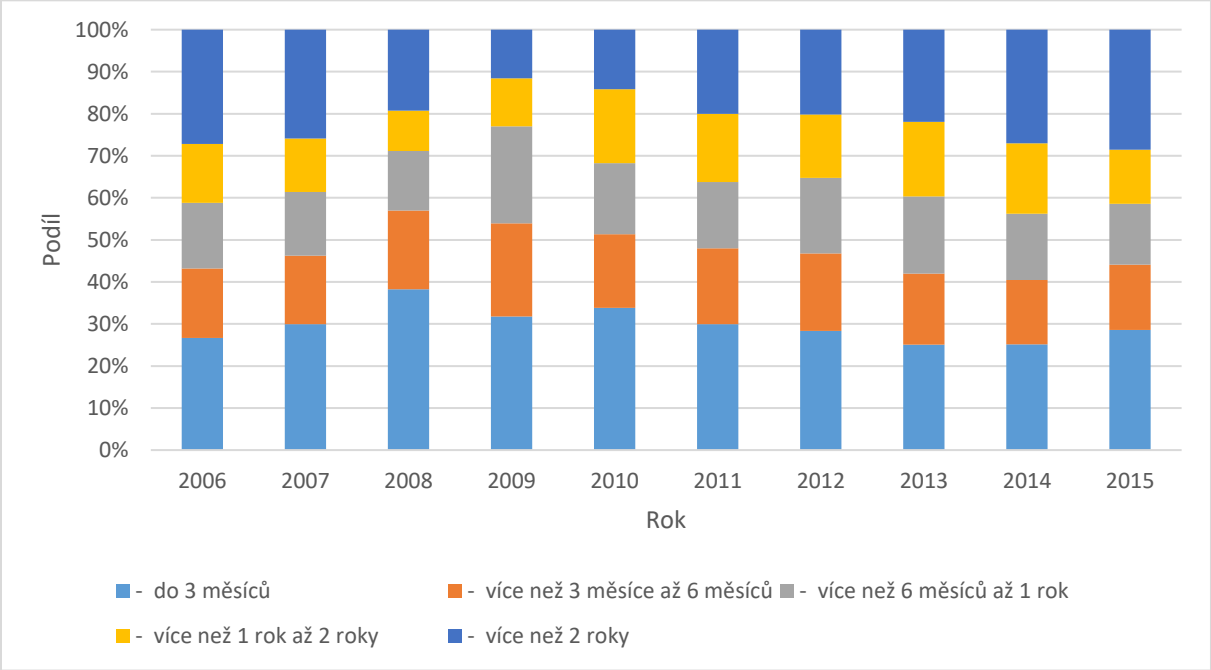
Oproti tomu data z MPSV mají tu výhodu, že se jedná o soubor všech nezaměstnaných, ale dle jiné definice. Na datech z MPSV by nebylo potřeba provádět statistické úsudky, pouze popsat soubor.

Důležitou otázkou je, zda jsou údaje z MPSV oproti údajům z VŠPS odlišné ve zkoumané veličině, tedy délce nezaměstnanosti. Jak ukazuji v kapitole 4.3.11, lidé registrovaní na Úřadě práce mají statisticky významně odlišnou délku nezaměstnanosti a nelze tato data brát za směrodatná ve vztahu k datům z VŠPS.

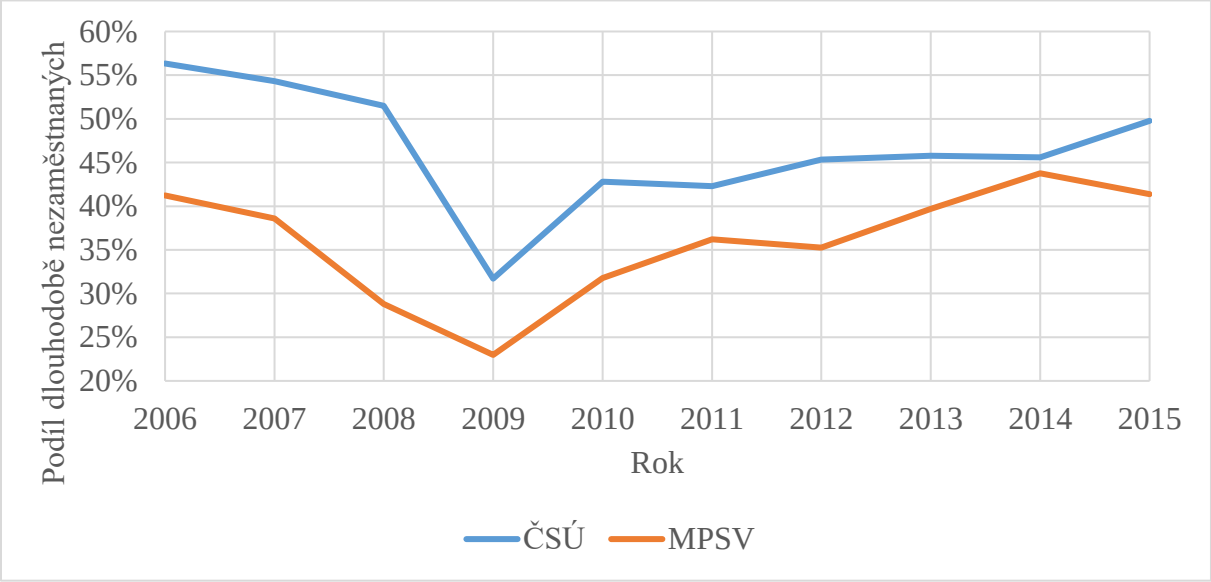
Tabulka 1.1 Srovnání rozdělení nezaměstnaných podle doby nezaměstnanosti v datech VŠPS a MPSV ve vybraných letech, zdroj: ČSÚ, MPSV

Doba nezaměstnanosti	2008 (VŠPS)	2008 (MPSV)	2010 (VŠPS)	2010 (MPSV)	2014 (VŠPS)	2014 (MPSV)
0 – 3 měsíce	16%	38%	18%	30%	20%	25%
3 – 6 měsíců	13%	19%	15%	18%	15%	15%
6 – 12 měsíců	19%	14%	24%	16%	19%	16%
12 – 24 měsíců	20%	10%	24%	16%	19%	17%
> 24 měsíců	32%	19%	19%	20%	26%	27%

Srovnání veřejně dostupných údajů ukazuje, že v datech z MPSV je vyšší podíl krátkodobě nezaměstnaných a nižší podíl dlouhodobě nezaměstnaných oproti údajům z VŠPS. V tabulce 1.1 je srovnání za roky, které používám v práci, obrázek 1.6 ukazuje graficky strukturu nezaměstnaných dle doby nezaměstnanosti stejně jako je v obrázku 1.2 a obrázek 1.7 srovnání podíly dlouhodobě nezaměstnaných v datech VŠPS a MPSV.



Obrázek 1.6 Roční rozdělení doby nezaměstnanosti v České republice, zdroj: MPSV



Obrázek 1.7 Roční podíl dlouhodobě nezaměstnaných v České republice, zdroj: ČSÚ, MPSV

1.4 Dopady a faktory nezaměstnanosti

Nezaměstnanost má dalekosáhlé dopady jak na jednotlivé nezaměstnané, tak přeneseně na celou společnost.

Z hlediska jedince jsou zřejmé finanční dopady, které můžou vést v extrémním případě až ke ztrátě bydlení. Nezaměstnanost může vést až k sebevraždě, dle odhadu uvedeného v Nordt a kol. (2015) se může jednat o zhruba 45 000 případů ročně, i obecněji k nárůstu rizika depresí, (Mossakowski, 2009).

Dopady sociální zahrnují fakt, že práce je výrobním faktorem, který je nezaměstnaností nedostatečně využíván. Dlouhodobá nezaměstnanost způsobuje ztrátu lidského kapitálu, kdy nezaměstnaní ztrácejí pracovní schopnosti a návyky. Nezaměstnaní lidé mají taktéž nižší střední dobu života (Anderton, 2006). Jiným sociálním dopadem je nárůst xenofobních nálad a protekcionismu, vyvolávaný strachem ze ztráty zaměstnání (Rotte a Steininger, 2008).

Nezaměstnanost nejvíce dopadá na mladé lidi a snižuje možnost jejich začlenění do společnosti. Rizikovými faktory jsou v tomto případě nízké vzdělání, pasivita na trhu práce, špatná podpora okolí i institucí, což diskutuje např. Kieselbach (2003).

Dlouhodobá nezaměstnanost je často spojována s dlouhodobým nárůstem obecné míry nezaměstnanosti, dle Layard a kol. (1991) dlouhodobě nezaměstnaní ztrácejí motivaci k dalšímu hledání práce a nevytvářejí tak tlak na snížení mezd. Jinou teorií vlivu dlouhodobé nezaměstnanosti na obecnou míru nezaměstnanosti dle Pissadires (1992) je vliv ztráty schopností dlouhodobě nezaměstnaných a jejich stále se zhoršující uplatnitelnost na trhu práce.

Jedním z obvykle zmiňovaných faktorů délky nezaměstnanosti je strnulost trhu práce. Například při srovnání portugalského a amerického trhu práce v Blanchard a Portugal (2001) byly míry nezaměstnanosti podobné, ale délka nezaměstnanosti byla v Portugalsku třikrát vyšší. Omezení na trhu práce vedlo jak k nižším mírám propouštění, tak nabírání nových zaměstnanců a jednotliví nezaměstnaní setrvávali bez práce déle.

Theodossiou a Zarotiadis (2010) na základě dat ze severního Řecka poukazují na efekt, podle kterého délka nezaměstnanosti negativně ovlivňuje dobu strávenou v dalším zaměstnání.

Jiným z faktorů působících na dobu nezaměstnanosti je efektivnější trh práce, který umožňuje rychlejší nalezení práce pro ty, kteří by ji tak jako tak našli relativně rychle. Oproti tomu déle nezaměstnaní pak zůstávají nezaměstnaní ještě o něco déle, jak diskutuje Portugal (2007).

Problémem může být i negativní motivace v podobě špatně nastaveného sociálního systému, která se diskutuje i v České republice, např. v Kotýnková (2006). Vyšší minimální mzda může být například jedním z důvodů déletrvající nezaměstnanosti a to především u těch, kteří mají na trhu práce obecně horší postavení, (Pedace a Rohn, 2011).

1.5 Doba nezaměstnanosti v předchozích výzkumech

Dobou nezaměstnanosti se zabývala a zabývá řada výzkumníků a to jak v zahraničí, tak v České republice, zde je uvádím.

Ze zahraničních výzkumů můžu jmenovat například práci Kerckhoffs a kol. (1994), kteří zkoumali dobu nezaměstnanosti v Nizozemí v 80. letech pomocí Coxova modelu. Cazes a Scarpetta (1998) analyzovali bulharská a polská individuální data a s jejich pomocí vliv ekonomické transformace na mechanismus propouštění a zaměstnávání a dobu nezaměstnanosti. Theodossiou a Yannopoulos (1998) se zabývali ve své práci rizikem propouštění pro specifické skupiny obyvatel v primárním a sekundárním sektoru ekonomiku a ukázali, že lidé z nižšího sektoru mají obecně delší dobu nezaměstnanosti. Mussida (2007) aplikoval model konkurenčních rizik (*competing risks model*) pomocí neparametrických metod a Weibullova rozdělení se závěrem, že ženatí muži mají nejkratší dobu nezaměstnanosti. Brick a Mlatsbeni (2007) v Kapském městě došli pomocí regresního modelu s Weibullovým rozdělením k závěru, že nejvyšší riziko dlouhodobé nezaměstnanosti mají ženy starající se o děti bez terciárního vzdělání a s omezenou znalostí angličtiny. Práce Kuhlenkasper a Steinhardt (2011) se zabývá modelováním doby nezaměstnanosti v Německu pomocí modelů s proměnlivým vlivem faktorů na rizikovou funkci v průběhu času se zvláštním důrazem na vliv pohlaví. Další prací je i Cuesta a Gonález (2014), kteří analyzovali vlivy na dobu nezaměstnanosti v Ekvádoru v letech 2007 – 2012 pomocí neparametrického modelu.

V České republice se tématem zabývali na počátku tohoto století Esser a Popelka (2003), kteří analyzovali data z ÚP v Příbrami pomocí Coxova modelu a pomocí modelu zrychleného času, Jarošová a kol. (2004) a Jarošová a Malá (2005), kteří analyzovali data z VŠPS pomocí modelu zrychleného času s log-logistickým rozdělením.

V poslední době vyšla řada článků a příspěvků na konferencích, zejména Malá, která se zabývá odhady intervalově cenzorovaných dat pomocí směsí rozdělení (2014) a pomocí L-momentů (2015) a Čabla (2012b, 2015 a 2016), který se zabývá modelováním doby nezaměstnanosti pomocí modelů zrychleného času.

Oproti těmto pracím se předkládaná disertační práce odlišuje především svou uceleností a zkoumáním rozdílů v době nezaměstnanosti v čase. V pracích kratšího rozsahu se jednotliví autoři věnují obvykle jedné zvolené metodice a menšímu počtu potenciálních vysvětlujících proměnných s jedním datovým přístupem (možné datové soubory viz kapitola 2).

Někteří autoři mají možnost pracovat s úplnými (necenzorovanými) daty a nepotřebují tak upravovat metodiku pro specifika intervalově cenzorovaných dat, která jsou obsažena ve VŠPS, tyto údaje ovšem obvykle nejsou mezinárodně srovnatelné. Druhým zásadním rozdílem je zkoumání na třech rozličných datových základech (viz kapitola 2). Novým přínosem je začlenění analýzy extrémních hodnot do kontextu analýzy doby nezaměstnanosti.

2 Data

V této kapitole je cílem představit data z VŠPS použitá v této práci. Údaje z VŠPS jsou publikovány agregovaně jako roční[3] nebo čtvrtletní[4] výsledky. Pro přístup k individuálním údajům je potřeba jednotlivé datové soubory nakoupit, buď za jedno čtvrtletí nebo za čtyři čtvrtletí dohromady[5]. Zdrojem některých údajů jsou pokyny v ČSÚ (2013).

V práci jsou rozlišena celkem tři období, každé zahrnuje celkem pět čtvrtletí. První období použité v práci začíná čtvrtým čtvrtletím roku 2007 a končí čtvrtým čtvrtletím roku 2008, toto období nazývám obdobím *před krizí*. Druhé začíná prvním čtvrtletím roku 2010, končí prvním čtvrtletím roku 2011 a nazývám jej *během krize*. Třetí období potom začíná čtvrtým čtvrtletím roku 2013 a končí čtvrtým čtvrtletím roku 2014, nazývám ho *po krizi*.

První a druhé období byla vybrána s ohledem na obecnou míru nezaměstnanosti v České republice – první období obsahuje čtvrtletí s nízkou mírou nezaměstnanosti (cca 4,3 – 5,2 %) druhé naopak s vysokou mírou nezaměstnanosti (cca 7,0 – 8,2 %). Třetí období bylo vybráno jako období klesající nezaměstnanosti, kdy se obecná míra pohybovala mezi výše uvedenými hodnotami (5,8 – 6,8 %). Vybraná období a vývoj míry nezaměstnanosti je zobrazen v obrázku 1.1.

Pro každé období byly vytvořeny tři specifické datové soubory. První z nich obsahuje všechny respondenty, kteří nalezli práci ve sledovaném období. Druhý z nich obsahuje všechny respondenty, kteří byli nezaměstnaní v průběhu sledovaného období, ať už práci nalezli, nebo nenalezli. Kromě těchto dvou souborů používám ještě třetí, který se skládá vždy ze čtvrtého čtvrtletí roků 2008, 2010 a 2014. V tomto třetím datovém souboru se striktně vzato nejedná o dobu do události ve smyslu analýzy přežití (viz kapitola 3.1), proto ji budu nazývat dobou nezaměstnanosti a nikoliv dobou do nalezení práce či zaměstnání. Vytvořené modely a úsudky v tomto třetím souboru jsou vytvářeny za předpokladu, že všichni nalezli práci v okamžiku šetření.

[3] na http://www.czso.cz/aktualni-produkt/41273

[4] na http://www.czso.cz/aktualni-produkt/41265

[5] ceník na https://www.czso.cz/csu/czso/cenik-informacnich-sluzeb-a-produktu-bwut?skupina=25

2.1 Doba nezaměstnanosti

Sledovanou veličinou je doba nezaměstnanosti. V každém šetření jsou šetřené osoby rozděleny do tří kategorií – zaměstnaní, nezaměstnaní a ekonomicky neaktivní dle definice ILO (viz kapitola 1.1). U zaměstnaných je zjišťováno, po jak dlouhou dobu jsou již zaměstnaní – označeno jako proměnná *ZamDob*. U nezaměstnaných je analogicky zjišťováno, po jak dlouhou dobu jsou nezaměstnaní – označeno jako proměnná *NezDob*. Doba nezaměstnanosti se určuje jako nižší z ukazatelů určujících jak dlouho osoba aktivně nepřerušeně hledá práci (*HledOd*) a jak dlouho je bez práce (*ExZamDat*). Za přerušení hledání práce se považuje neaktivita na trhu práce delší než 4 týdny. Obě tyto proměnné jsou intervalově cenzorované (více o intervalovém cezorování v kapitole 3.2), hodnoty shrnuje tabulka 2.1.

Tabulka 2.1 Doba zaměstnanosti a nezaměstnanosti v měsících v datech VŠPS – kódování

NezDob	Doba nezaměstnanosti	ZamDob	Doba zaměstnanosti
1	$\langle 0;1\rangle$	1	$\langle 0;1\rangle$
2	$\langle 1;3\rangle$	2	$\langle 1;3\rangle$
3	$\langle 3;6\rangle$	3	$\langle 3;6\rangle$
4	$\langle 6;12\rangle$	4	$\langle 6;12\rangle$
5	$\langle 12;18\rangle$	5	$\langle 12;36\rangle$
6	$\langle 18;24\rangle$	6	$\langle 36;60\rangle$
7	$\langle 24;48\rangle$	7	$\langle 60;120\rangle$
8	$\rangle 48$	8	$\langle 120;240\rangle$
		9	$\rangle 240$

U první datového souboru, který obsahuje všechny dotazované, kteří nalezli práci v průběhu šetření, je potřeba určit dobu do nalezení práce, která je tvořena intervalem, ve kterém L označuje spodní hranici intervalu, R označuje horní hranici intervalu a $Doba2\check{S}$ označuje dobu uplynulou mezi dvěma po sobě jdoucími šetřeními. Hranice získáme z rovnice

$$\begin{aligned} L &= \max(NezDob\_L + Doba2\check{S} - ZamDob\_R;0,01) \\ R &= NezDob\_R + Doba2\check{S} - ZamDob\_L \end{aligned} . \qquad (2.1)$$

Tedy k době nezaměstnanosti v šetření, kdy byl dotyčný ještě nezaměstnaný, přičteme dobu mezi 2 po sobě jdoucími šetřeními (*Doba2Š*) a odečteme dobu, po kterou již je v práci v dalším šetření. Neměla by nastat situace, že doba v novém zaměstnání je vyšší než doba mezi dvěma šetřeními, tedy $ZamDob \leq 3$. Celkem u osmnácti šetřených osob byla tato skutečnost zjištěna, z datového souboru byly odstraněny. Za povšimnutí stojí, že levý interval může při výpočtu vyjít záporně – taková hodnota je potom převedena na 0,01 (některá uvažovaná parametrická rozdělení pro popis a modelování vyžadují kladné hodnoty náhodné veličiny, proto nepoužívám jako minimální hodnotu 0).

U druhého datového souboru byly k osobám z prvního přidány všechny osoby, které byly nezaměstnané a práci si v průběhu šetření nenalezly. U těchto osob je doba do nalezení práce vyšší než hodnota *NezDob_L* v posledním šetření, ve kterém se vyskytly. Jedná se o tzv. zleva cenzorované pozorování (pojem je opět vysvětlen v kapitole 3.2).

Tabulka 2.2 obsahuje četnosti všech tří užitých přístupů u všech tří období. Pokud bychom se zaměřili na podíl těch, kteří nalezli v průběhu šetření práci, k těm, co byli nezaměstnaní, zjistíme, že se mírně zvýšil z přibližně 27 % na 30 %, respektive 31 %. Tento ukazatel nicméně nemá jednoznačnou interpretaci, protože neznáme přesný časový rámec, ke kterému bychom tento podíl přiřadili.

Tabulka 2.2 Četnosti v jednotlivých obdobích a datových souborech

	Název	Před krizí	Během krize	Po krizi
Nalezli práci	Soubor 1	624	1069	964
Všichni nezaměstnaní	Soubor 2	2318	3558	3116
Vybrané čtvrtletí	Soubor 3	1233	1894	1478

2.2 Vysvětlující proměnné

Ve VŠPS se sleduje celá řada vybraných proměnných, které můžeme použít jako vysvětlující proměnné pro dobu nezaměstnanosti. Níže poskytuji jejich přehled.

- *Regiony soudržnosti* (*region*) vyjadřuje, ve kterém regionu soudržnosti (správní členění NUTS2) se zkoumaná domácnost nachází. Tato proměnná byla zjištěna z okresu domácnosti.

- *Počet osob v domácnosti* (*pocod*) udává celkový počet osob ve zkoumané domácnosti respondenta.
- *Vztah k osobě v čele domácnosti* (*vzocd*) popisuje, zda je zkoumaná osoba v čele hospodařící domácnosti – v úplné rodinné domácnosti je to vždy muž bez ohledu na ekonomickou aktivitu dotyčného či ostatních členů domácnosti.
    - Pokud je v domácnosti více rodinných domácností, určuje se osoba v jejím čele podle ekonomické aktivity.
    - Je-li hospodařící domácnost tvořena neúplnou rodinnou domácností, upřednostňuje se ekonomicky aktivní osoba v pořadí rodič, dítě, prarodič. Jestliže ani jedna osoba není ekonomicky aktivní, pak se zase postupuje ve stejné posloupnosti, tedy přednost má rodič.
    - Je-li hospodařící domácnost tvořena nerodinnou domácností, rozhoduje ekonomická aktivita.
    - Ostatní osoby kromě osoby v čele domácnosti se rozdělují do kategorií *Manžel/ka či Životní partner*/*ka*, *Dítě* a *Ostatní,* kam spadají rodiče a další příbuzní nebo členové domácnosti.
- *Pohlaví* (*pohl*) určuje pohlaví zkoumané osoby.
- *Rodinný stav* (*rodstav*) popisuje oficiální rodinný stav, tedy dle stavu uvedeného v občanském průkazu. Možnosti jsou *Svobodný/á*, *Ženatý*/*Vdaná, Ovdovělý*/*á* a *Rozvedený*/*á*.
- *Věk* zachycuje věk dotyčné osoby. Kromě samotného věku se určuje *Věková skupina* (*veksk*), která je tvořena pětiletými kategoriemi počínaje kategorií 15 – 19.
- *Země narození* (*zemnar*) kóduje, ve které zemi se dotyčný narodil. V práci užívám kategorie *ČR*, *Slovensko, Ukrajina* a *Jiná*, přičemž občas jsou druhá a třetí kategorie přiřazeny do poslední na základě nedostatečného počtu pozorování.
- *Státní příslušnost* (*stprisl*) určuje státní příslušnost dotyčného. Kategorie odpovídají kategoriím proměnné *zemnar*.
- *Zdravotní postižení* (*zdrpost*) je zachycením statusu osoby, které byl připsán status osoby se zdravotním postižením s 1., 2. nebo 3. stupněm invalidity. Z originálních tří kategorií používám dvě – *Ne* a *Ano*.

- *Vzdělání* (*ISCED*) je určením vzdělání zkoumané osoby. Používá se zde zjednodušená stupnice, ve které 1 značí *Bez vzdělání*, 2 je *Základní vzdělání včetně neukončeného*, 3 označuje *Středoškolské vzdělání bez maturity*, 4 pak *Středoškolské vzdělání s maturitou* a 5 je kategorie *Vysokoškolského vzdělání*.
- *Registrace na Úřadu práce* (*regup*) je proměnnou určující, zda je dotyčná osoba registrovaná na Úřadu práce či ne.
- *Pracovní zkušenost* (*exzam*) popisuje existenci pracovní zkušenosti. *Ano* označuje člověka, který již práci měl, *Ne* označuje člověka, který práci ještě neměl.
- *Typ hledaného úvazku* (*hleduv*) určuje preferovaný typ úvazku. Možnosti jsou *Na plnou pracovní dobu, Na plnou pracovní dobu/zkrácenou, Na zkrácenou pracovní dobu/plnou, Na zkrácenou pracovní dobu* a *Jakýkoliv úvazek*.
- *Metody hledání* představují souhrn metod, které uchazeči používají či nepoužívají pro nalezení práce. Na otázky se odpovídá *Ano* či *Ne*.
    - *Metoda A* (*hledmeta*) značí, že si osoba hledá práci pomocí ÚP.
    - *Metoda B* (*hledmetb*) označuje, že si osoba hledá práci pomocí soukromých zprostředkovatelů.
    - *Metoda C* (*hledmetc*) říká, že osoba přímo kontaktuje zaměstnavatele.
    - *Metoda D* (*hledmetd*) označuje, že si osoba hledá práci prostřednictvím příbuzných či známých.
    - *Metoda E* (*hledmete*) značí, že osoba hledá práci aktivně pomocí inzerátů.
    - *Metoda F* (*hledmetf*) značí, že osoba hledá práci pasivně pomocí inzerátů.
    - *Metoda G* (*hledmetg*) značí, že se osoba účastní pohovorů.
    - *Metoda J* (*hledmetj*) označuje, že osoba očekává výsledek vyřízení žádosti o práci.
    - *Metoda K* (*hledmetk*) označuje osobu, která očekává zprávu z ÚP.
    - *Metoda L* (*hledmetl*) říká, že osoba očekává výsledek konkurzu na zaměstnání ve veřejném sektoru.
    - *Metoda M* (*hledmetm*) je odpovědí na otázku, zda si dotyčný hledá práci jiným způsobem.

- *Počet užitých metod hledání zaměstnání* (*hleda*) je počet odpovědí *Ano* na otázky *hledmeta, hledmetb, hledmetc, hledmetd, hledmete, hledmetf, hledmetg* a *hledmetm,* což označuje aktivní způsoby hledání práce. Hodnoty jsou potom celočíselné na škále 0 až 8.
- *Velikost obce* (*velikostobce*) je proměnná určující počet obyvatel obce, ve které se nachází zkoumaná domácnost. Je to kategoriální proměnná s kategoriemu *Do 2000, 2000 – 4999, 5000 – 9999* a *10 000 a více*.

Kromě těchto proměnných obsahují datové soubory i váhy jednotlivých osob (viz kapitola 1.3). Konkrétní hodnoty těchto proměnných a jejich četnosti jsou uvedeny v Příloze C.

3 Metodologie

Cílem této kapitoly je popsat metody užité v práci při zkoumání doby do nalezení práce. V první podkapitole rozebírám základy analýzy přežití, v dalších šesti pak jednotlivé metody z analýzy přežití, přičemž kde je to možné, uvádím i alternativy. V osmé podkapitole diskutuji postup vytváření modelů, v předposlední stručně představuji teorii extrémní hodnoty a model po částech spojené distribuční funkce, v poslední uvádím použitý software.

Úvodními zdroji pro tuto kapitolu jsou monografie a učebnice zaměřené na analýzu přežití – Klein a Moeschberger (1997), Lawless (2003), Kleinbum a Klein (2012) a Liu (2012), dále monografie věnující se tématice intervalových dat – Sun (2006) a Chen a kol. (2013) a monografie zaměřené na regresní modely včetně modelů užívaných v rámci analýzy přežití – Harrell jr. (2001) a Hosmer a kol. (2008).

3.1 Analýza přežití

Analýza přežití je obecný pojem zastřešující soubor metod užívaných pro zkoumání doby do nastání určité události. Touto událostí může být např. úmrtí, vyléčení, porucha, či nalezení práce a podle typu události se analýza přežití aplikuje v celé škále vědních oborů, namátkou třeba v

- medicíně, kde se užívá k hodnocení klinických pokusů zkoumajících nové léky, léčebné postupy či zařízení,
- biologii, kde se užívá k modelování evolučních procesů,
- sociálních vědách, ve kterých se můžou zkoumat nezaměstnanost, užívání drog, či manželství,
- demografii, kde se kromě zkoumání samotné úmrtnosti zkoumají např. počátek užívání antikoncepce, migrace či věk při prvním porodu,
- inženýrství, ve kterém se analýza přežití používá ke zkoumání doby životnosti.

Analýza přežití vznikla původně jako obor určený ke zkoumání mortality na základě matričních záznamů. Nejranější analýzy můžeme vystopovat do 17. století do práce anglického statistika Johna Graunta, který publikoval první známé úmrtnostní tabulky v roce 1662. (Graunt, 1939)

Ve druhé polovině 20. století nastal velký rozvoj analýzy přežití díky Coxovu modelu a jím vytvořené metodě částečné věrohodnosti (Cox, 1972). Nejnovějším polem rozvoje analýzy přežití je systém čítacích procesů a teorie martingálů.

Základními cíli analýzy přežití jsou obvykle odhad a interpretace funkce přežití a/nebo rizikové funkce, porovnání různých funkcí přežití a vyjádření vztahu mezi vysvětlujícími proměnnými a pravděpodobnostním rozdělením času do události.

V analýze přežití tedy zkoumáme spojitou nezápornou náhodnou veličinu *T*. Základním popisem jejího pravděpodobnostního rozdělení je pravděpodobnost, že tato doba bude větší než čas *t*, funkce přežití je tedy definována jako

$$S(t) = P(T > \mathrm{t}) = 1 - F(t) = \int_t^{\infty} f(s)ds. \tag{3.1}$$

Funkce přežití je doplňkem distribuční funkce, z čehož vyplývají její vlastnosti:

- je nerostoucí,
- v čase $t = 0$ je hodnota $S(0) = 1$ a $\lim_{t\to\infty}(S(t)) = 0$,
- je zleva spojitá.

Dalším způsobem popisu pravděpodobnostního rozdělení v analýze přežití je riziková funkce *h*(*t*), pro kterou se užívají i jiné názvy, například podmíněná míra selhání v analýze spolehlivosti, míra úmrtnosti v demografii, věkem podmíněná míra selhání v epidemiologii či převrácená hodnota Millovy míry v ekonomii. Tato funkce je okamžitým potenciálem nastání události za jednotku času za předpokladu, že tato událost ještě nenastala, vyjádřeno vzorcem tedy je

$$h(t) = \lim_{\Delta t\to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}. \tag{3.2}$$

Riziková funkce *h*(*t*) je nezáporná a v jakémkoliv okamžiku *t* může nabývat libovolné hodnoty od nuly po nekonečno. Její průběh může být libovolný a odráží vlastnosti sledovaného jevu. Základní tvary této funkce se věnovali např. Klein a Moeschberger (1997):

- Rostoucí, který se vyskytuje všude tam, kde pravděpodobnost nastání sledované události roste s časem, nejčastěji tedy v případech sledování stárnutí a opotřebovávání strojů.
- Klesající tvar se vyskytuje naopak tam, kde pravděpodobnost nastání sledované události s časem klesá, typicky při sledování pooperačních stavů, kde sledovanou událostí je smrt pacienta např. po transplantaci.
- Vanovitý tvar (*bathtub-shaped*) mají funkce, kde je nejdříve klesající pravděpodobnost nastání sledované situace a následně rostoucí. Jedná se o častý průběh této pravděpodobnosti při sledování úmrtnosti lidí od narození, kdy je nejdříve vysoké riziko

úmrtnosti, které rychle klesne, aby po zbytek života rostlo. Dalším obdobným případem je uvádění některých složitějších zařízení do provozu, které mají zpočátku vysokou míru poruchovosti např. z důvodu chybných, předem netestovatelných, součástek, která po naběhnutí do provozu klesne a začne se zvyšovat zase s rostoucí mírou opotřebení.

- Kopcovitý tvar (*hump-shaped*) má potom taková funkce, při níž pravděpodobnost nastání sledované události nejdříve roste k určitému maximu a následně klesá. Tento tvar bývá typický například pro sledování úmrtnosti po úspěšných chirurgických zákrocích, kdy riziko úmrtí kulminuje až s určitým odstupem od operace, nebo dobu do rozvodu od sňatku.

Mezi rizikovou funkcí $h(t)$, funkcí přežití $S(t)$ a hustotou pravděpodobnosti je jednoznačný vztah (Klein, Moeschberger, 1997)

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}, \tag{3.3}$$

ze kterého plyne i rovnice

$$S(t) = \exp\left[-\int_0^t h(s)ds \right]. \tag{3.4}$$

Rovnice (3.5) ukazuje výpočet kumulované rizikové funkce $H(t)$, která se občas používá pro neparametrický Nelsonův-Aalenův odhad, navržený v Nelson (1972) a matematicky formalizovaný v Aalen (1978), přičemž mezi $S(t)$ a $H(t)$ je jednoznačný vztah vyplývající z rovnic (3.4) a (3.5).

$$H(t) = \int_0^t h(s)ds, \tag{3.5}$$

$$S(t) = \exp\left[-H(t) \right]. \tag{3.6}$$

3.2 Cenzorování

Cenzorování je proces neodmyslitelně spojený s analýzou přežití. Jedná se o jakýkoliv postup vedoucí k cenzorovanému pozorování, přičemž některé příklady uvádím níže. Zde je na místě odlišit analýzu přežití od analýzy cenzorovaných dat, protože zcela obecně může hodnota cenzorované proměnné X nabývat libovolné hodnoty, čímž se liší od proměnné T definované výše. V praxi i literatuře jsou oba pojmy obvykle volně zaměňovány, protože málokdy dochází k výskytu cenzorovaných hodnot mimo rámec analýzy přežití. V dalších odstavcích budu dodržovat značení náhodné veličiny doby do výskytu události T.

Obecně můžeme říct, že hodnota libovolného pozorování leží v intervalu (L, R], budeme tedy předpokládat

$$T \in (L, R], \tag{3.7}$$

kde $L \leq R$. Můžeme rozlišit následující čtyři situace:

- Pokud platí $L = R$, jedná se o úplné pozorování s hodnotou R.
- Pokud platí $L = 0$, jedná se o zleva cenzorované pozorování v intervalu $(0, R]$.
- Pokud platí $R = \infty$, jedná se o zprava cenzorované pozorování v intervalu (L, ∞).
- Pokud neplatí ani jedno z výše zmíněných, pak se jedná o intervalově cenzorované pozorování v intervalu $(L, R]$.

Cenzorovaná pozorování můžou vznikat z mnoha příčin.

Zleva cenzorované pozorování může například nastat, pokud podstoupí člověk test na přítomnost určité nemoci a ten je pozitivní – doba od nakažení nemocí není přesně známa, jenom víme, že je člověk již infikován.

Zprava cenzorované pozorování může nastat například tak, že je člověk sledován po určitou dobu v rámci studie a událost u něj nenastane do okamžiku ukončení studie či vyřazení člověka ze studie. Známe tak pouze minimální dobu do nastání dané události. Tento typ cenzorování je v praxi nejčastější, především v medicíně a sledování životnosti výrobků.

Intervalové cenzorování může nastat například tak, že událost nastane mezi dvěma kontrolami, ale nevíme přesně kdy. V této práci nastává událost – nalezení práce, mezi dvěma pohovory s pracovníkem ČSÚ.

V analýze přežití se obvykle předpokládají tři základní vlastnosti vztahu zkoumané náhodné veličiny T a době cenzorování C.

1. Nezávislost – Rizikové funkce cenzorovaných a necenzorovaných pozorování jsou stejné.
2. Náhodnost – Rizikové funkce cenzorovaných a necenzorovaných pozorování jsou stejné pro všechny pozorované podskupiny.
3. Neinformativnost – Rozdělení doby do nastání události T neposkytuje žádnou informaci o rozdělení doby cenzorování C a obráceně.

Všechny níže uvedené metody analýzy přežití předpokládají pro nezkreslené odhady neinformativní cenzorování (Klein, Moeschberger, 1984).

Intervalově cenzorovaná pozorování můžeme rozdělit na čtyři typy – I. typu, II. typu, dvojitě cenzorovaná a panelová data (*panel count data*). První dva jsou převzaty z Groeneboom

a Wellner (1992), třetí z De Gruttola a Lagakos (1989) a poslední z Kalbfleisch a Lawless (1985).

3.2.1 Intervalové cenzorování I. typu

Pod tímto typem se označují taková data, která vždy obsahují buď nulu, nebo nekonečno, tedy ve kterých jsou jednotlivá pozorování buď cenzorována zleva, nebo zprava. Tento typ intervalového cenzorování se vyskytuje tehdy, když je každý subjekt studie pozorován pouze jednou a zjišťuje se, zda sledovaná událost již nastala nebo ještě nenastala. V prvním případě je to tedy cenzorování zleva, ve druhém zprava. Tento typ dat obvykle pochází z průřezových studií.

Data tohoto typu intervalového cenzorování jsou obvykle reprezentována formou

$$\{C, \delta = I(T \leq C)\}. \tag{3.8}$$

kde C je čas pozorování, I je indikátorová funkce a proměnná δ značí, zda je pozorování cenzorované zleva nebo zprava.

3.2.2 Intervalové cenzorování II. typu

Jedná se o intervalové cenzorování, kdy se v pozorovaném souboru vyskytuje alespoň jeden interval obsahující obě hodnoty L a R, neobsahuje tedy hodnoty $L = 0$ ani $R = \infty$. Data tohoto typu jsou občas prezentována formou

$$\{W, V, \delta_1 = I(T \leq W), \delta_2 = I(W < T \leq V), \delta_3 = 1 - \delta_1 - \delta_2\}, \tag{3.9}$$

kde se předpokládá, že každá jednotka je pozorována dvakrát v časech W a V, přičemž platí $W \leq V$. Pokud položíme $W = V = C$, můžeme použít vzorec 3.9 pro popis intervalového cenzorování I. typu.

3.2.3 Dvojité cenzorování

Předpokládejme existenci dvou propojených událostí s časem výskytu X a S, pro které platí $X \leq S$. Zajímá nás doba mezi těmito událostmi $T = S - X$. Za dvojitě cenzorovaná pozorování budeme považovat taková pozorování, pro které platí, že obě doby X a S jsou intervalově cenzorované, $X \in (L, R]$ a $S \in (W, V]$, kde $L \leq R$ a $W \leq V$.

Tento typ dat v sobě zahrnuje jako speciální případy intervalové cenzorování II. typu, pokud známe přesnou hodnotu X ($L = R$), a cenzorování zprava, pokud známe přesnou hodnotu S ($W = V$) nebo pokud je S cenzorovaná zprava.

3.2.4 Panelová data

Panelová data jsou data generována opakovaným pozorováním stejných jednotek v průběhu času a zaznamenáváním stavu dané jednotky, tedy víme, zda daná událost (změna stavu) nastala mezi dvěma pozorováními. V tomto typu dat může často sledovaná událost nastat vícekrát.

3.3 Turnbullův odhad

Statistické úsudky se v modelech analýzy přežití obvykle provádí pomocí věrohodnostní funkce. Příspěvek i-tého necenzorovaného pozorování k věrohodnosti je roven hodnotě hustoty pravděpodobnosti v bodě pozorování, příspěvek cenzorovaného pozorování je pak za podmínky neinformativnosti na základě diskuze v Kiefer a Wolfowitz (1956) dle Peto (1973) funkcí vektoru parametrů $\boldsymbol{\theta}$ a je roven integrálu z hustoty pravděpodobnosti přes interval cenzorování $(L_i, R_i]$, tedy

$$L_i(\boldsymbol{\theta}) = \int_{L_i}^{R_i} f(t,\boldsymbol{\theta})dt = F(R_i,\boldsymbol{\theta}) - F(L_i,\boldsymbol{\theta}) = S(L_i,\boldsymbol{\theta}) - S(R_i,\boldsymbol{\theta}). \tag{3.10}$$

Věrohodnostní funkce $L(\boldsymbol{\theta})$ je součinem příspěvků jednotlivých pozorování podle typu cenzorování. Z rovnice (3.9) lze dovodit, že příspěvek zprava cenzorovaného pozorování k věrohodnosti je roven $S(L_i)$ a příspěvek zleva cenzorovaného pozorování je roven $1 - S(R_i)$.

Kromě věrohodností funkce se často pracuje s jejím logaritmem $l(\boldsymbol{\theta}) = \ln[L(\boldsymbol{\theta})]$.

V této kapitole uvažuji maximálně věrohodný neparametrický odhady (*NPMLE*). Pokud se v datech vyskytují mimo necenzorovaných již pouze zprava cenzorovaná pozorování, bývají používány jako maximálně věrohodné odhady odhad Kaplanův-Meierův (1958) nebo odhad Nelsonův-Aalenův (Nelson, (1972), Aalen, (1978)), přičemž druhý z nich je odhadem kumulativní rizikové funkce (viz rovnice (3.5)).

Kaplanův-Meierův ani Nelsonův-Aalenův odhad nejsou definovány po posledním pozorování, pokud je toto zprava cenzorované, protože NPMLE není v této oblasti jednoznačný. To znamená, že průběh funkce přežití po tomto bodu již neovlivňuje věrohodnost odhadu. Střední hodnota a další charakteristiky se obvykle odhadují za předpokladu, že nejvyšší hodnota se rovná cenzorované hodnotě (Efron, 1967), což ovšem vede pro malé soubory

k negativně zkresleným odhadům (Klein, 1991). Alternativou je předpokládat pevně zvolenou hodnotu (Gill, 1980). Takovýto odhad střední hodnoty se označuje jako omezený průměr (*restricted mean – rmean*)

Tento problém s neurčitostí nastává i u intervalově cenzorovaných dat, což se odráží i v podobě výsledného odhadu, který má podobu rozdělení pravděpodobnosti, že sledovaná událost nastala v určitém časovém intervalu, ale s neznámým rozdělením pravděpodobnosti uvnitř tohoto intervalu. Existuje několik možností, jak odhadnout NPMLE, v práci je použit Turnbullův odhad (1974, 1976) s úpravou dle Gentleman a Greyer (1994) pro naplnění Kuhn-Tuckerových podmínek.

Turnbullův odhad je iterativním EM algoritmem, který nachází řešení simultánních rovnic

$$S_j = \pi_j\left(S_{1,}, S_2, \ldots S_n\right) \quad (1 \le j \le m). \tag{3.11}$$

EM algoritmy obecně střídají kroky očekávání E(*expectation*), ve kterém je vytvořena funkce očekávaní logaritmu věrohodnosti při současné hodnotě parametrů, a kroky maximalizace M(*maximization*), ve kterém se určují hodnoty parametrů maximalizující věrohodnostní funkci vytvořenou v kroku E, (Dempster a kol., 1977).

Klein a Moeschberger (1997) popisují Turnbullův odhad takto:

Nechť je řada bodů v čase $\tau_0 < \tau_1 < \ldots < \tau_m$ obsahující všechny body L_i, R_i, pro pozorování $i = 1, 2, \ldots, n$. Pro i-té pozorování definujeme váhu α_{ij}, která nabývá hodnoty 1, pokud interval $(\tau_{j-1}, \tau_j]$ je obsažen v intervalu (L_i, R_i), a 0 jinak. α_{ij} je indikátorem toho, zda událost, která nastala v intervalu (L_i, R_i), mohla nastat v čase τ_j. Takto vzniká prvotní odhad $S(\tau)$. Následuje algoritmus:

1. Spočítejme pravděpodobnost, že událost mohla nastat v čase τ_j

$$p_j = S(\tau_{j-1}) - S(\tau_j), \quad j = 1, 2, \ldots, m. \tag{3.12}$$

2. Odhadněme počet událostí, které nastaly v čase τ_j

$$d_j = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_{k=j}^{m} \alpha_{ij} p_k} \quad j = 1, 2, \ldots, m. \tag{3.13}$$

3. Spočítejme odhadovaný počet jednotek v riziku v čase τ_j

$$Y_j = \sum_{k=j}^{m} d_k. \tag{3.14}$$

4. Spočítejme aktualizovaný odhad funkce přežití $S(t)$ podle odhadů z kroků 2 a 3. Pokud je aktualizovaný odhad podobný předchozímu, končíme iterativní proces. Pokud ne, opakujeme kroky 1 až 3 za použití aktualizovaného odhadu.

Konvergence tohoto EM algoritmu sama o sobě nezaručuje nalezení globálního maxima věrohodnostní funkce, ale pokud splní Kuhnovy-Tuckerovy podmínky, tak se jedná o NPMLE. Tyto podmínky jsou:

$$1-\sum_{j=1}^{m} p_j = 0, \tag{3.15}$$

$$p_j \geq 0 \quad j = 1, 2, ..., m, \tag{3.16}$$

$$u_j p_j = 0 \quad j = 1, 2, ..., m, \tag{3.17}$$

$$u_j \geq 0 \quad j = 1, 2, ..., m, \tag{3.18}$$

$$\frac{\partial}{\partial p_j}\left\{ l(p) + \sum_{j=1}^{m} p_j \left(u_j - u_0\right)\right\} = d_j + u_j - u_0 = 0 \quad j = 1, 2, ..., m. \tag{3.19}$$

První dvě podmínky definují odhadované pravděpodobnosti, další tři podmínky se týkají Lagrangeových multiplikátorů u_j, (Gentleman a Greyer, 1994).

Vedle Turnbullova odhadu existují i další odhadovací metody pro NPMLE, z nichž dvě nejznámější jsou ICM (Iterative Convex Minorant) algoritmus diskutovaný v Jonglobed (1998) a směs těchto dvou přístupů, EM-ICM algoritmus navržený ve Wellner a Zhan (1997)

Na základě NPMLE se odhadují kvantily $x_p = 1 - \hat{S}^{-1}(p)$ a omezený průměr (*restricted mean*), což je průměr vypočtený z funkce přežití omezené shora pěvně zvolenou konstantou. Při výpočtu průměru se předpokládá, že funkce přežití je v každém z odhadnutých intervalů lineárně klesající. Volba konstanty pak ovlivňuje hodnotu takového průměru prodloužením či zkrácením posledního intervalu.

3.4 Vážený log-rank test

Cílem tohoto testu je obecně testovat hypotézu o rovnosti distribučních funkcí, tedy

$$H_0 : S_1(t) = S_2(t) = ... = S_k(t). \tag{3.20}$$

Pro necenzorovaná data i data zprava cenzorovaná existuje řada neparametrických testů testujících shodu distribučních funkcí, tedy i funkcí přežití. Pro data intervalově cenzorovaná je potřeba tyto testy dále upravit. Tohoto typu testu se využívá obvykle k testování, zda se

funkce přežití liší podle hodnot vybrané vysvětlující proměnné. Zde představuji obecný postup testování založený na sdruženém kontinuálním modelu (*grouped continuous model*), viz Fay (1996, 1999).

Seřadíme potenciální časy do události podle velikosti $0 = \tau_0 < \tau_1 < \ldots < \tau_m < \tau_{m+1} = \infty$. Výběrový prostor rozdělíme na $m + 1$ intervalů, s j-tým značeným $I_j = (\tau_{j-1}, \tau_j]$. Předpokládáme, že všechny body L_i a R_i jsou obsaženy v časech τ_j.

$$\alpha_{ij} = I\left(L_i < \tau_j < R_i\right). \tag{3.21}$$

Obecný model funkce přežití pro i-tou jednotku je

$$P\left(T_i > \tau_j \middle| \mathbf{x_i}\right) = S\left(\tau_j \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right), \tag{3.22}$$

kde $\mathbf{x_i}$ je vektor vysvětlujících proměnných, $\boldsymbol{\beta}$ je vektor parametrů k těmto proměnným a $\boldsymbol{\gamma}$ je vektor parametrů neznámé funkce přežití. Tato funkce přežití se může modelovat různými způsoby, zde užíváme model s urychleným selháním (AFT model), který více popisuji v kapitole 3.5. Základem je vztah

$$\ln\left(T\right) = \mu + \mathbf{x}^T \boldsymbol{\beta} + \sigma\varepsilon, \tag{3.23}$$

kde známe rozdělení chybové složky ε. Rovnici 3.22 upravíme pomocí monotónní transformace g, která je popsána pro všechny časové okamžiky τ_j a i-tou jednotku pomocí vektoru parametrů $\boldsymbol{\gamma}$.

$$S\left(\tau_j \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right) = 1 - F\left\{g\left(T_i\right) - \mathbf{x}_i^T \boldsymbol{\beta}\right\}. \tag{3.24}$$

Věrohodnostní funkce sdruženého kontinuálního modelu pro intervalově cenzorovaná data má podle rovnice (3.10) podobu

$$L = \prod_{i=1}^{n} \sum_{j=1}^{m+1} \alpha_{ij} \left[S\left(\tau_{j-1} \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right) - S\left(\tau_j \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right] = \prod_{i=1}^{n} \left[S\left(L_i \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right) - S\left(R_i \middle| \mathbf{x}_i^T \boldsymbol{\beta}, \boldsymbol{\gamma}\right)\right] \tag{3.25}$$

Derivací $\ln(L)$ podle vektoru $\boldsymbol{\beta}$ získáme pro $\boldsymbol{\beta} = \mathbf{0}$ skórovou statistiku U. Maximálně věrohodný odhad vektoru $\boldsymbol{\gamma}$ pro $\boldsymbol{\beta} = \mathbf{0}$ je Turnbullův odhad popsaný v předchozí kapitole. Vektor skórů pro parametr vektor parametrů $\boldsymbol{\beta}$ můžeme zapsat jako

$$\mathbf{U} = \sum_{i=1}^{n} \mathbf{x_i} \left(\frac{\hat{S}'\left(L_i\right) - \hat{S}'\left(R_i\right)}{\hat{S}\left(L_i\right) - \hat{S}\left(R_i\right)} \right) = \sum_{i=1}^{n} \mathbf{x_i} c_i, \tag{3.26}$$

kde $\hat{S}'(t)$ je hodnota derivace $S(t)$ podle $\boldsymbol{\beta}$ při $\boldsymbol{\beta} = \mathbf{0}$. Pro l-tý řádek můžeme psát

$$U_l = \sum_{j=1}^{m} w_j \left[d_{jl}^* - \frac{n_{jl}^* d_j^*}{n_j^*} \right], \tag{3.27}$$

kde d_{jl}^* je očekávaná hodnota počtu událostí v I_j pro l-tou skupinu při platnosti nulové hypotézy, d_j^* je očekávaná hodnota celkového počtu událostí v I_j při platnosti nulové hypotézy a podobně n_{jl}^* a n_j^* jsou očekávané počty pozorování.

Vektor skórů **U** tak lze považovat za vektor hodnot, které popisují vážený součet rozdílů mezi očekávanou a skutečnou hodnotou počtu událostí pro jednotlivé funkce přežití, obvykle tedy rozdělené podle hodnot vysvětlující proměnné. Tento vektor má asymptoticky normální rozdělení s nulovým vektorem středních hodnot a kovarianční maticí **V**.

Podle konkrétních hodnot přiřazených vektorům **w** a $\mathbf{c_i}$ můžeme vytvořit různé testové statistiky. V této práci užívám vážený log-rank test podle Suna (1996), kde

$$c_i = \frac{\hat{S}(L_i)\ln\left(\hat{S}(L_i)\right) - \hat{S}(R_i)\ln\left(\hat{S}(R_i)\right)}{\hat{S}(L_i) - \hat{S}(R_i)} \tag{3.28}$$

a $w_j = 1$. Vzhledem k jednotkové váze pro všechny hodnoty *j* popisuje tedy vektor skórů rozdíl mezi očekávanou a skutečnou hodnotou počtu událostí. Jednoduchou transformací dostáváme testové kritérium Q, které má při platnosti nulové hypotézy asymptoticky Chí-kvadrát rozdělení o $k - 1$ stupních volnosti, kde k je počet testovaných funkcí přežití, tedy počet hodnot vysvětlující proměnné

$$Q = \mathbf{U}^T \mathbf{V}^{-1} \mathbf{U} \approx \chi^2(k-1). \tag{3.29}$$

Můžeme uvažovat alternativní hypotézu ve tvaru

$$H_1 : non\, H_0, \tag{3.30}$$

nebo ve tvaru

$$H_1 : S_1(t) > S_2(t) > \ldots > S_3(t), \tag{3.31}$$

$$H_1 : S_1(t) < S_2(t) < \ldots < S_3(t). \tag{3.32}$$

Pokud uvažujeme alternativní hypotézu (3.31) nebo (3.32), pak se obvykle hovoří o testu trendu.

Pro test trendu předpokládejme vektor konstant **k** obsahující hodnoty vysvětlující kvantitativní proměnné. Platí, že

$$\mathbf{k}^T \mathbf{U} \approx N\left(0, \mathbf{k}^T \mathbf{V} \mathbf{k}\right) \tag{3.33}$$

a tedy

$$Z = \frac{\mathbf{k}^T \mathbf{U}}{\sqrt{\mathbf{k}^T \mathbf{V} \mathbf{k}}} \approx N(0,1). \tag{3.34}$$

Z je pak testové kritérium pro test trendového vztahu mezi hodnotami vysvětlující proměnné a funkcí přežití.

Obdobný model s jinými hodnotami $\mathbf{c_i}$ a $\mathbf{w}$ byl navržen Finkelsteinem (1986). Výhoda Sunova modelu je, že se jedná o zobecnění nejběžněji používaného log-rank testu pro zprava cenzorovaná data a dá se tak pro ně použít. Jinou často používanou možností je použít zobecnění Wilcoxonova-Mannova-Whitneyova testu, navrženého Peto a Peto (1972).

3.5 Model zrychleného času

Kromě neparametrických metod se v analýze přežití užívají i různé parametrické regresní modely, nejčastěji se jedná o plně parametrický model proporčního rizika (*proportional hazards model* - PH), Coxův (1972) semi-parametrický model proporčního rizika a model zrychleného času (*Accelerated Failure Time model* – AFT model).

Modely proporčního rizika jsou modely, ve kterých se předpokládá multiplikativní působení vysvětlujících proměnných na základní rizikovou funkci $h_0(t)$. V plně parametrickém modelu je základní riziková funkce specifikována pomocí známého pravděpodobnostního rozdělení, zatímco v Coxově modelu je základní riziková funkce neparametrická. Nevýhodou plně parametrického modelu je omezení na dvě pravděpodobnostní rozdělení – exponenciální a Weibullovo, jejichž riziková funkce je konstantní resp. monotónní. Tyto modely se odhadují metodou maximální věrohodnosti, popsanou ve Finkelstein (1986).

Coxův model je flexibilnější, neboť $h_0(t)$ je neparametrická, ale základní metodou odhadu je metoda částečné věrohodnosti, která není aplikovatelná na intervalově cenzorovaná data a nevede k nezkresleným odhadům. Existují různé návrhy jiných metod odhadu, např. metoda marginální věrohodnosti diskutovaná v Satten (1996) či Goggins a kol. (1998). Betensky a spol. (2002) navrhují užít metodu lokální věrohodnosti a Cai a Betensky (2003) navrhují metodu založenou na penalizovaných splinech.

AFT model je regresním modelem, který předpokládá závislost pravděpodobnostního rozdělení logaritmu času na poloze μ, měřítku σ, vektoru vysvětlujících proměnných $\mathbf{x_i}$ pro i-tou jednotku, vektoru parametrů $\boldsymbol{\beta}$ a chybové složce ε, u níž známe rozdělení, ve tvaru

$$\ln(T_i) = \mu + \mathbf{x}_i^T \boldsymbol{\beta} + \sigma\varepsilon. \tag{3.35}$$

Vyjádříme funkci přežití a upravíme

$$S_i(t) = P\left[\left(\mu + \mathbf{x}_i^T \boldsymbol{\beta} + \sigma\varepsilon\right) \geq \ln(t)\right] = P\left(\varepsilon \geq \frac{\ln(t) - \mu - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right). \tag{3.36}$$

Funkci přežití pak můžeme rozdělit na základní funkci přežití S_0 a složku popisující vliv vektoru vysvětlujících proměnných

$$S(t|\mathbf{x}) = S_0\left(\frac{\ln(t) - \mu - \mathbf{x}^T \boldsymbol{\beta}}{\sigma}\right), \tag{3.37}$$

kde S_0 je funkce přežití chybové složky ε a $\mathbf{x}^T\boldsymbol{\beta}$ určující posun polohy T se nazývá akcelerační faktor. Rovnice (3.35) až (3.36) můžeme také vyjádřit jen pro čas do události místo jeho logaritmu

$$T = \exp(\mu)\exp\left(\mathbf{x}^T \boldsymbol{\beta}\right)\exp(\sigma\varepsilon), \tag{3.38}$$

$$S_i(t) = P\left(\exp(\mu)\exp\left(\mathbf{x_i^T}\boldsymbol{\beta}\right)\exp(\sigma\varepsilon) \geq t\right) = S_0(t)\left[t\exp\left(\mathbf{x_i^T}\boldsymbol{\beta}\right)\right], \tag{3.39}$$

$$S_0(t) = P\left[\exp(\mu + \sigma\varepsilon \geq t)\right]. \tag{3.40}$$

Z rovnice (3.39) je vidět vztah vektoru vysvětlujících proměnných k funkci přežití. Jestliže výraz $\exp(\mathbf{x}^T\boldsymbol{\beta})>1$, působí vysvětlující proměnné celkově zpomalujícím způsobem – ke sledovaným událostem dochází v průměru později, a obráceně.

Vzhledem k jednoznačnému vztahu mezi funkcí přežití $S(t)$ a rizikovou funkcí $h(t)$ lze dle rovnice (3.4) odvodit AFT model i z pohledu rizikové funkce

$$h(t|\mathbf{x}) = \frac{1}{\sigma t} h_0\left(\frac{\ln(t) - \mu - \mathbf{x^T}\boldsymbol{\beta}}{\sigma}\right) = h_0(t)\left[t\exp\left(-\mathbf{x^T}\boldsymbol{\beta}\right)\right]\exp\left(-x^T\beta\right), \tag{3.41}$$

kde $h_0(t)$ je rizikovou funkcí ε.

3.6 Úsudky v rámci modelu zrychleného času

Věrohodnostní funkce příslušná AFT modelu je při užití rovnice (3.10)

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) = \prod_{i=1}^{n} S\left[\left(\frac{\ln(L_i) - \mu - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right) - S\left(\frac{\ln(R_i) - \mu - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}\right)\right] \tag{3.42}$$

a její logaritmus je roven

$$\ln\left(L(\boldsymbol{\theta})\right) = \sum_{i=1}^{n} \ln\left[S\left(\frac{\ln(L_i) - \mu - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right) - S\left(\frac{\ln(R_i) - \mu - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right) \right]. \tag{3.43}$$

Pro každé *i*-té pozorování lze získat vektor skórových statistik jako vektoru parciálních derivací podle vektoru parametrů $\boldsymbol{\theta}$

$$\mathbf{U_i}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln\left(L_i(\boldsymbol{\theta})\right) \tag{3.44}$$

a pomocí druhých derivací získáme kovarianční matici

$$\mathbf{V_i}(\boldsymbol{\theta}) = -\left(E \frac{\partial^2 L_i}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right). \tag{3.45}$$

Pokud platí podmínka neinformativnosti cenzorování (kapitola 3.2), uplatňuje se centrální limitní věta na vektor součtů, tedy celková skórová statistika

$$\mathbf{U}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{U}_i(\boldsymbol{\theta}) \tag{3.46}$$

má asymptoticky normální rozdělení s nulovým vektorem středních hodnot a kovarianční maticí

$$\mathbf{V}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{V}_i(\boldsymbol{\theta}). \tag{3.47}$$

Z rovnosti

$$\mathbf{U}(\boldsymbol{\theta}) = \mathbf{0} \tag{3.48}$$

získáváme vektor odhadnutých parametrů $\hat{\boldsymbol{\theta}}$, pro který platí

$$\hat{\boldsymbol{\theta}} \approx N\left[\mathbf{0}, \mathbf{V}(\boldsymbol{\theta})^{-1}\right] \tag{3.49}$$

Pro testování hypotéz o parametrech z vektoru $\boldsymbol{\theta}$ využijeme Fisherovu informační matici, jejíž střední hodnota je obsažena v matici $\mathbf{V}(\boldsymbol{\theta})$

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\left(\frac{\partial^2 \ln\left(L(\hat{\boldsymbol{\theta}})\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right). \tag{3.50}$$

Odhad kovarianční matice odhadu parametrů je matice inverzní k matici **I**. Odmocniny prvků na hlavní diagonále odhadu kovarianční matice poskytují odhady směrodatných chyb odhadů parametrů. (Liu, 2012)

Testujeme-li hypotézu $H_0: \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$, můžeme využít Waldovu statistiku (Wald, 1943)

$$\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{T}\mathbf{I}\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right), \tag{3.51}$$

která má za platnosti nulové hypotézy asymptoticky Chí-kvadrát rozdělení. Použijeme-li odmocninu z této statistiky a přepíšeme obecně pro jeden parametr β, dostaneme testovou statistiku

$$Z=\frac{\hat{\beta}-\beta_0}{s.e.\left(\hat{\beta}\right)}\approx N(0;1), \tag{3.52}$$

kde *s.e.* značí směrodatnou chybu odhadu parametru β a β_0 je hodnota parametru β dle nulové hypotézy. Tato testová statistika se používá pro testy jednotlivých parametrů.

Druhý důležitý test v rámci úsudků nad věrohodnostní funkcí je test věrohodnostním poměrem. Nulová hypotéza tohoto testu je H_0: $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$ pro všechny parametry v $\hat{\boldsymbol{\theta}}$ nebo $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ pro část vektoru $\hat{\boldsymbol{\theta}}$. Testová statistika má podobu

$$\Lambda=2\ln L\left(\hat{\boldsymbol{\theta}}\right)-2\ln L(\boldsymbol{\theta})\approx\chi^{2}(M) \tag{3.53}$$

a asymptoticky chí-kvadrát rozdělení s *M* stupni volnosti, kde $L(\boldsymbol{\theta})$ je věrohodnostní funkce části modelu bez jednoho nebo více parametrů, $L\left(\hat{\boldsymbol{\theta}}\right)$ je věrohodnostní funkce modelu se všemi parametry a *M* je rozdíl v počtu parametrů.

Kromě testování statistické významnosti rozdílu mezi různými modely s různými počty proměnných je možné též modely řadit na základě vhodného kritéria založeného na věrohodnosti odhadu. V práci pro porovnávání modelů používám Akaikeho informační kritérium (*AIC*, Akaike (1973, 1974)) a jako doplňkovou informaci uvádím i hodnoty Bayesovského informačního kritéria (*BIC*, Schwarz (1978)). Jestli *p* je počet parametrů modelu, *L* je hodnota věrohodnosti a *n* je počet pozorování, pak

$$AIC=2p-2\ln(L), \tag{3.54}$$

$$BIC=p\ln(n)-2\ln(L). \tag{3.55}$$

3.7 Log-logistické rozdělení

V rámci AFT modelu lze použít různá rozdělení, pokud mají nebo je lze převézt na rozdělení charakterizovaná parametry polohy a měřítka tak, aby odpovídala specifikaci v rovnici (3.34). V angličtině se tato rozdělení označují jako *location-scale* a patří sem například normální

rozdělení, log-normální, Weibullovo, exponenciální, Cauchyho, logistické, log-logistické, Laplaceovo či Studentovo.

V této práci jsem zvolil log-logistické rozdělení. Hlavním kritériem byla hodnota věrohodnostní funkce (viz kapitola 4.1), která byla nejvyšší vždy u tohoto nebo u log-normálního rozdělení. Obě rozdělení dávaly téměř totožné odhady průběhu funkce přežití, z log-logistického rozdělení lze navíc získat i interpretaci v podobě poměru šancí na nastání sledované události.

Jestliže chybová složka ε z regresního modelu (3.34) má logistické rozdělení, potom doba do události T sleduje log-logistické rozdělení. Standardní parametrizace základních funkcí a charakteristik log-logistického rozdělení pro náhodnou veličinu T, kde $\alpha > 0$ a $\beta > 0$ je

$$S(t) = \frac{1}{1+(t/\alpha)^{\beta}}, \tag{3.56}$$

$$h(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1+(t/\alpha)^{\beta}}, \tag{3.57}$$

$$E(X) = \frac{\alpha\pi/\beta}{\sin(\pi/\beta)}, \quad \text{pokud}\, \beta > 1, \tag{3.58}$$

$$\text{Modus} = \alpha\left(\frac{\beta-1}{\beta+1}\right)^{1/\beta}, \quad \text{pokud}\ \beta > 0, \tag{3.59}$$

$$D(X) = \alpha^2\left(2\frac{\pi/\beta}{\sin(2\pi/\beta)} - \frac{(\pi/\beta)^2}{\sin^2(\pi/\beta)}\right), \quad \text{pokud}\, \beta > 2, \tag{3.60}$$

$$t_p = \alpha\left(\frac{p}{1-p}\right)^{1/\beta}, \tag{3.61}$$

přičemž ve vztahu k AFT modelu platí

$$\mu = \ln(\alpha) \quad \text{a} \quad \sigma = 1/\beta. \tag{3.62}$$

Z rovnic (3.37) a (3.62) je vidět, že akcelerační faktor $\exp(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$ přímo násobí hodnotu parametru α, viz

$$T = \exp(\mu)\exp(\mathbf{x}^T\boldsymbol{\beta})\exp(\sigma\varepsilon) = \exp(\ln(\alpha))\exp(\mathbf{x}^T\boldsymbol{\beta})\exp(\sigma\varepsilon) = \alpha\exp(\mathbf{x}^T\boldsymbol{\beta})\exp(\sigma\varepsilon). \tag{3.63}$$

Z toho také vyplývá, že akcelerační faktor působí přímo na kvantilovou funkci (3.61), hodnoty větší než 1 zvyšují hodnoty kvantilů doby do události, tedy zpomalují plynutí času tak, jak bylo napsáno obecně v kapitole 3.5.

Z výše uvedených rovnic (3.56) – (3.61) také vyplývají některé vlastnosti log-logistického rozdělení. Funkce rizika je monotónní klesající, pokud je $\beta \leq 1$, jinak je nejprve rostoucí a poté klesající. Střední hodnota není definovaná, pokud je $\beta \leq 1$, a rozptyl není definovaný, pokud je $\beta \leq 2$.

Jak bylo uvedeno výše, AFT model s log-logistickým rozdělením umožňuje interpretaci pomocí poměru šancí (*odds ratio* - *OR*). Poměr šancí je poměrem pravděpodobnosti, že událost nastane po okamžiku *t* a pravděpodobnosti, že nastane do okamžikem *t*. Píšeme

$$OR\left[S\left(t;\mathbf{x},\boldsymbol{\beta}^*,\alpha,\beta\right)\right]=\frac{S\left(t;\mathbf{x},\boldsymbol{\beta}^*,\alpha,\beta\right)}{1-S\left(t;\mathbf{x},\boldsymbol{\beta}^*,\alpha,\beta\right)}=\frac{\left[1+\alpha\exp\left(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*\right)t^{\beta}\right]^{-1}}{\left[\alpha\exp\left(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*\right)t^{\beta}\right]^{-1}}, \tag{3.64}$$

nahradíme $\left[\exp(\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta})^*\right]^{-1}=\exp\left(-\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*\right)$ a můžeme upravit

$$OR\left[S\left(t;\mathbf{x},\boldsymbol{\beta}^*,\alpha,\beta\right)\right]=\exp\left(-\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*\right)\frac{\left[1+\alpha t^{\beta}\right]^{-1}}{\left[\alpha t^{\beta}\right]^{-1}}=\exp\left(-\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*\right)\frac{S\left(t|\mathbf{x}=\mathbf{0}\right)}{1-S\left(t|\mathbf{x}=\mathbf{0}\right)}. \tag{3.65}$$

Poměr šancí na přežití je pro daný vektor kovariant **x** dán výrazem exp(-$\mathbf{x}^{\mathbf{T}}\boldsymbol{\beta}^*$), efekt jednotkové změny proměnné x_r na šanci přežití je dán výrazem exp($-\beta_r^*$). Ve vztahu k AFT modelu pak platí $\beta^*=-\beta/\sigma$, tedy vliv jednotkové změny proměnné x_r na šanci přežití je dán výrazem exp(β_r/σ).

Log-logistické rozdělení je jediné známé, které umožňuje interpretaci jak ve smyslu zrychleného času, tak ve smyslu poměru šancí na přežití.

3.8 Budování regresního modelu

Při budování vícerozměrného regresního modelu se mezi sebou střetávají dva základní principy – jeden z nich říká, že model má obsahovat všechny podstatné vysvětlující proměnné. Podle druhého z nich má být model co nejjednodušší, tedy obsahovat minimální množství proměnných. V této práci používám datově orientovaný přístup, ve kterém zkoumám postupně jednotlivě všechny potenciální vysvětlující proměnné, poté tvořím úplný model obsahující všechny tyto proměnné najednou a z něho pak odvozuji minimální adekvátní model.

Pokud je vysvětlující proměnná v modelu s jednou proměnnou kvantitativní, rozhoduji o vhodnosti tohoto modelu Waldovým testem na hladině významnosti 0,05. Pokud je tato vysvětlující proměnná nominální, slučuji jednotlivé kategorie tak dlouho, dokud model s menším počtem kategorií nemá statisticky významně nižší věrohodnost podle testu

věrohodnostním poměrem při hladině významnosti 0,05, poté slučování zastavuji. Pro ordinální kategorie postupuji podobně, zde však slučuji vždy dvě sousedící kategorie.

Pokud například obsahuje nominální proměnná kategorie *A, B, C*, pak můžu sloučit dvojice *A* a *B, B* a *C* či *A* a *C*. Pokud ordinální proměnná obsahuje kategorie 1, 2, 3, pak slučuji vždy dvě sousedící kategorie, tedy 1 a 2 či 2 a 3, ale ne dvojici kategorií 1 a 3. Toto zachovávám i tehdy, pokud nejsou nesousedící kategorie statisticky významně odlišné především z důvodu grafického vyjádření vztahu doby do nalezení práce a vysvětlující proměnné.

Pokud mohu vytvořit s jednou vysvětlující proměnnou dva modely – jeden využívající tuto proměnnou jako kvantitativní a jeden jako ordinální kategoriální, volím ten model, který má statisticky významně vyšší věrohodnost při testu věrohodnostním poměrem na hladině významnosti 0,05. Pokud je výsledek testu statisticky nevýznamný, dávám přednost modelu s kvantitativní vysvětlující proměnnou.

Tímto postupem dostávám zjednodušené modely, přičemž sloučení jednotlivých kategorií považuji za příznak nedostatečné odlišnosti v době nezaměstnanosti. V těchto modelech zkoumám nejen vliv jednotlivých proměnných, ale i změnu tohoto vlivu v čase.

Úplný model je označení pro model se všemi vysvětlujícími proměnnými.

Minimální adekvátní model je označení pro model, ze kterého byly vyřazeny všechny statisticky nevýznamné vysvětlující proměnné či sloučeny statisticky nevýznamně odlišné kategorie v rámci vysvětlující proměnné. Kroky pro tvorbu tohoto modelu jsou následující:

1. Začínám s úplným modelem. Některé zdroje doporučují zařazovat všechny proměnné, které v jednotlivém zkoumání měly hladinu významnosti příslušného testu nižší než 0,20 či 0,25 a proměnné klinicky (na základě určité teorie) významné, (Bendel a Afifi (1977), Costanza a Afifi (1979), Mickey a Greenland, (1989)).
2. Následně slučuji kategorie proměnných, které mají p-hodnotu Waldova testu vyšší než 0,1. Testuji model se sloučenými kategoriemi proti předcházejícím modelu s touto kategorií zpětně i pomocí testu věrohodnostním poměrem opět na hladině významnosti 0,1.
3. Zkoumám, zda vyřazení proměnné nezměnilo *významně* odhady ostatních parametrů. Za *významné* zde v souladu s Hosmer a kol. (2008) považuji změny větší než 20 % hodnoty odhadu a změny, které mění hodnocení statistické významnosti Waldova testu. Takto pokračuji až do vyřazení všech vyhovujících proměnných a sloučení všech vyhovujících kategorií.

4. Zpětně zkouším zařadit jednotlivé vyřazené proměnné do modelu. Pokud je změna věrohodnosti statisticky významná na hladině významnosti 0,1, proměnná v modelu zůstává.

Obecně mají postupy založené na tvorbě modelu postupnými kroky (*stepwise*) tyto nevýhody:

- Obvyklé testové statistiky nemají očekávaná rozdělení, (Grambsch a O´Brien, 1991).
- Výsledné odhady směrodatných chyb jsou zkresleně nízké, což vede k užším intervalům spolehlivosti a nižším p-hodnotám, (Altman a Andersen, 1989), (Derksen a Keselman, 1992)
- Regresní koeficienty jsou zkresleně vysoké. (Chatfield, 1991)
- Volba zahrnutých regresních koeficientů závisí nejen na jejich skutečných hodnotách, ale i na hodnotách jejich odhadů. (Copas a Long, 1991)

Tyto nevýhody jsou v mém postupu částečně akceptovány a adresovány následujícími rozhodnutími:

- Postupuji zpětně od úplného modelu. To mi umožňuje posoudit statistickou významnost s nezkreslenými odhady směrodatných chyb, které jsou v úplném modelu.
- Zpětný postup umožňuje lépe pracovat s multikolinearitou (Mantel, 1970).
- Pracuji s hladinou významnosti 0,1 pro jednotlivé proměnné, tedy s vyšší než obecně používanou hladinou významnosti.

V práci na výše popsaný postup navazuji zařazováním interakcí proměnných. Interakce je pojem značící zařazení násobku dvou (či více) proměnných do modelu jako samostatné proměnné, jejíž hodnoty je možné standardně odhadovat a testovat. Postup je následující:

1. Zkusím zařazení interakcí všech statisticky významných proměnných a zjišťuji p-hodnotu testu věrohodnostním poměrem mezi modelem bez interakce a s interakcí.
2. Seřadím interakce podle p-hodnoty a postupně je zařazuji do modelu. Pokud dojde k vylepšení modelu na hladině významnosti 0,01, nechávám interakci v modelu a zkouším další. Pokud nedojde k vylepšení modelu, interakci nezařazuji a zkouším další.
3. Poté, co vyzkouším všechny interakce, u kterých p-hodnota v prvním kroku byla nižší než 0,01, končím.

Zařazení interakcí umožňuje na jednu stranu model zpřesnit, na druhou stranu se komplikuje interpretace modelu. Pro ilustraci uvažujme lineární regresní model se závislou

proměnnou proměnnou Y a se dvěma dichotomickými nezávislými proměnnými X_1 a X_2. Model s interakcí pak můžeme zapsat

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{3.66}$$

Můžeme dovodit, že

- parametr β_1 udává změnu hodnoty Y, pokud je proměnná X_2 rovna 0,
- parametr β_2 udává změnu hodnoty Y, pokud je proměnná X_1 rovna 0,
- pokud je proměnná X_2 rovna 1, pak změna hodnoty X_1 z 0 na 1 se projeví na hodnotě Y jako změna o součet hodnot parametrů β_1 a β_2,
- pokud je proměnná X_1 rovna 1, pak změna hodnoty X_2 z 0 na 1 se projeví na hodnotě Y jako změna o součet hodnot parametrů β_1 a β_2.

Ve vícerozměrném modelu s více interakcemi, jaký prezentuji v kapitole 4.5, je důležité být opatrný s interpretací, protože změna závislé proměnné v reakci na změnu nezávislé proměnné nemusí být na první pohled zřejmá a může dojít ke zkreslení vztahu závislé proměnné ke zkoumané dvojici nezávislých proměnných, pokud hrají roli ještě další proměnné. Hladina významnosti byla zvolena 0,01 právě z toho důvodu, aby se modely nestaly příliš komplikované.

V grafických výstupech prezentujících interakce (viz Příloha) jsou uvedeny přepočítané koeficienty vlivu celých kombinací oproti základním hodnotám. Jelikož jsou odhady parametrů normálně rozdělené náhodné veličiny se známými středními hodnotami, rozptyly a kovariancemi, lze vliv konkrétní kombinace dopočítat jako součet náhodných veličin.

3.9 Teorie extrémní hodnoty

Teorie extrémních hodnot je odvětvím matematické statistiky, které se zabývá zaprvé stochastickým chováním výběrového maxima nebo minima nezávislých, stejně rozdělených náhodných veličin. Zadruhé pak studuje stochastické chování nezávislých, stejně rozdělených náhodných veličin, jejichž hodnota překračuje určitou dostatečně vysokou hodnotu *u*. (Vojtěch, 2011)

Praktické aplikace teorie extrémních hodnot zahrnují různé oblasti, např. meteorologie (maxima a minima teplot či srážek), oceánské inženýrství (nejvyšší vlny v oblasti), hydrologie (maxima říčních průtoků) či pojišťovnictví (maximální plnění z přírodních katastrof).

3.9.1 Rozdělení výběrového maxima

V teorii extrémních hodnot se uvažuje rozdělení výběrového maxima, kde se uplatňuje Fisherova-Tippettova-Gněděnkova věta (občas se uvádí jako Fisherova-Tippettova věta). Podle Fréchet (1927), (Fisher a Tippett, 1928) a Gněděnko (1943):

Mějme posloupnost nezávislých, stejně rozdělených náhodných veličin X_1, X_2,…, X_q. Definujme posloupnost částečných maxim jako

$M_1 = X_1,$

$M_2 = \max(X_1, X_2)$

.

$M_q = \max(X_1, X_2, \ldots, X_q)$

Nechť X_1, X_2,…,X_q je posloupnost nezávislých, stejně rozdělených náhodných veličin s nedegenerovanou distribuční funkcí F. Pokud existují posloupnosti normalizačních konstant $\sigma_q > 0$ a $\mu_q \in R$ (závislých na q) a nějaká nedegenerovaná distribuční funkce G taková, že platí

$$\lim_{q\to\infty} P\left[\frac{M_q - \mu_q}{\sigma_q} \le z\right] = \lim_{q\to\infty} P\left(M_q \le \mu_q + \sigma_q z\right) = \lim_{q\to\infty}\left[F(\mu_q + \sigma_q z)\right]^q = G(z), \qquad (3.67)$$

pak distribuční funkce $G(z)$ musí být jednoho ze tří typ rozdělení a to pouze:

$$\text{Fréchetovo:} \quad \Phi_\alpha(z) = \begin{cases} 0, & z \le 0, \\ \exp\left(-z^{-\alpha}\right), & z > 0,\ \alpha > 0, \end{cases}$$

$$\text{Weibullovo:} \quad \Psi_\alpha(z) = \begin{cases} \exp\left(-(-z)^{\alpha}\right), & z \le 0, \alpha > 0, \\ 1, & z > 0, \end{cases} \qquad (3.68)$$

$$\text{Gumbellovo:} \quad \Lambda(z) = \exp(-\exp(-z)), \qquad x \in R.$$

Tato tři rozdělení se souhrnně nazývají Fisherova-Tippettova rozdělení nebo taky rozdělení extrémních hodnot. Tato rozdělení mohou být zpětně upravena zařazením parametrů polohy μ a měřítka σ

$$\textit{Fréchetovo:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \le 0, \\ \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^{-\alpha}\right), & x > 0,\ \alpha > 0, \end{cases}$$

$$\textit{Weibullovo:} \quad \Psi_\alpha(x) = \begin{cases} \exp\left(-\left(-\frac{x-\mu}{\sigma}\right)^{\alpha}\right), & x \le 0, \alpha > 0, \\ 1, & x > 0, \end{cases} \qquad (3.69)$$

Gumbellovo: $\Lambda(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\sigma}\right)\right), \qquad x \in R.$

Vhodnou transformací (Gněděnko, 1943) pak získáváme distribuční funkci

$$G(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right]\right\}. \tag{3.70}$$

Rozdělení popsané distribuční funkcí G(x) se nazývá zobecněným rozdělením extrémních hodnot (Generalized Extreme Value Distribution – *GEV*). Platí, že pokud $\xi > 0$, odpovídá *GEV* Fréchetovu rozdělení, pokud $\xi < 0$, odpovídá *GEV* Weibullovu rozdělení a pokud $\xi = 0$, tak pro limitu $\xi \to 0$ získáme Gumbelovo rozdělení.

Zobecněným rozdělením extrémních hodnot tedy můžeme odhadovat maxima nezávislých náhodných výběrů, což se v praxi uplatňuje v tzv. metodě blokových maxim (*block maxima method*), která se nejčastěji používá pro modelování pravděpodobnosti maxim či minim v časových řadách. Příkladem je modelování maximálních ročních říčních průtoků ve Vojtěch (2010). Tato metoda spočívá v rozdělení posloupnosti pozorování do stejně velkých bloků, přičemž se modelují maxima v těchto blocích. U říčních průtoků se obvykle rozdělí hodnoty po jednotlivých letech a pracuje se tak s ročními maximy.

3.9.2 Rozdělení špiček nad prahem

Metoda blokových maxim může opomenout relevantní pozorování, jejichž využití by mohlo přinést přesnější odhady, diskuze např. v článku Ferreira a De Haan (2015) nebo Jarušková a Hanek (2006). V některých případech mohou být v jednom bloku vedle maxima další hodnoty, které jsou považovány za extrémní a v jiném bloku by třeba byly maximem.

Proto se často užívá analýza rozdělení špiček nad prahem (*peaks over threshold – POT*), ve které se pracuje se všemi hodnotami přesahujícími určitý, dostatečně vysoký, práh (*threshold*) u. Obvykle uvažujeme hodnoty $X - u$, tedy rozdíly mezi hodnotou náhodné veličiny X a prahem u. Podle de Haan a Balkema (1974) a Pickands (1975) platí Pickandova-Balkemova-de Haanova věta:

Uvažujme nezávislé, stejně rozdělené náhodné veličiny X s obecně neznámou distribuční funkcí F. Náhodné veličiny, pro které platí $X > u$, mají podmíněnou přesahovou (*excess*) distribuční funkci

$$F_u(y) = P(X - u \le y \,|\, X > u) \text{ pro } 0 \le \varpi_F - u, \tag{3.71}$$

kde u je daný práh, $y = x - u$ a jsou hodnoty pozorování nad prahem a ω_F je pravý koncový bod distribuční funkce F. Pro velkou třídu distribučních funkcí platí

$$F_u(y) \to Z_{\xi,\mu,\sigma} = 1 - \left(1 + \xi \frac{x - \mu^*}{\sigma}\right)^{-\frac{1}{\xi}} \qquad \text{pro } u \to \infty. \tag{3.72}$$

Parametr μ^* zde představuje tzv. dolní mez nosiče (support) zobecněného Paretova rozdělení (*GPD*), pod který daná funkce nikdy neklesne, přičemž se často uvažuje jeho nulová hodnota. Z je distribuční funkce zobecněného Paretova rozdělení, které můžeme opět rozdělit na tři rozdělení podle hodnoty parametru ξ, pro $\xi > 0$

$$\textit{Paretovo:} \qquad Z(x) = 1 - \left(\frac{x - \mu^*}{\sigma}\right)^{-\xi}, \qquad x > \sigma, \tag{3.73}$$

pro $\xi < 0$

$$\textit{Beta rozdělení:} \qquad Z(x) = 1 - \left(-\frac{x - \mu^*}{\sigma}\right)^{\xi}, \qquad -\sigma \le x \le 0, \tag{3.74}$$

pro $\xi = 0$

$$\textit{Exponenciální:} \qquad Z(x) = 1 - \exp\left(-\frac{x - \mu^*}{\sigma}\right), \qquad x \ge 0. \tag{3.75}$$

Stěžejní roli v obou metodách hraje parametr ξ, který odráží chování konce rozdělení náhodné veličiny X. Tento parametr se občas nazývá index extrémních hodnot. Pokud konec rozdělení klesá mocninou funkcí, je rozdělení výběrových maxim Fréchetovo. Do této skupiny patří rozdělení s těžkými konci, např. Paretovo, Cauchyho či studentovo. Pokud je konec rozdělení konečný, je rozdělení výběrových maxim Weibullovo. Sem patří např. rovnoměrné nebo beta rozdělení. Pokud konec klesá exponenciálně, je rozdělení výběrových maxim limitně Gumbelovo. Do této skupiny patří rozdělení s tzv. štíhlými konci, např. normální, logaritmicko-normální, exponenciální či gama rozdělení. Pro přehlednost je vše uvedeno v tabulce 3.1.

Tabulka 3.1 Srovnání *GEV* a *GPD*

ξ	*GEV*	*GPD*	Příklady rozdělení
> 0	Fréchetovo	Paretovo	Paretovo, Cauchyho, Studentovo
< 0	Weibullovo	Beta	rovnoměrné, beta
= 0	Gumbelovo	Exponenciální	normální, logaritmicko-normální, exponenciální, gama

V metodě špiček nad prahem jsou důležitá dvě rozhodnutí. Zaprvé určení hodnoty prahu u a zadruhé určení metody odhadu. Určení hodnoty prahu u se obvykle dělá pomocí ad hoc grafických postupů, jejich diskuze viz např. Tanaka a Takara (2002) nebo ve vztahu k modelu po částech spojené distribuční funkce Scarrott a MacDonald (2012). Z těchto postupů pro intervalově cenzorovaná data lze použít graf stability odhadu parametru ξ pro různé hodnoty prahu nebo stability odhadu střední hodnoty nad prahem – tato metoda je adaptací zobrazení stability výběrového průměr nad prahem (sample mean excess). Kromě toho se zobrazuje i průběh statistiky vhodného testu dobré shody s empirickou distribuční funkcí, např. Kolmogorovův-Smirnovův. Tuto nicméně pro intervalově cenzorovaná data neznáme, řešila by se tedy shoda s Turnbullovým odhadem. V práci místo tohoto ukazatele používám logaritmus hodnoty věrohodnostní funkce.

Tyto metody vychází ze dvou úvah – na jedné straně čím vyšší bude zvolený práh, tím spíše bude rozdělení pravého konce konvergovat ke zobecněnému Paretovu rozdělení. Na druhé straně čím nižší bude zvolený práh, tím více pozorování se použije k odhadu parametrů GPD a jejich odhad bude tedy přesnější.

Pro odhady parametrů zobecněného Paretova rozdělení se používají různé metody, z nichž nejvýznamnější jsou metoda maximální věrohodnosti, diskutovaná v Smith (1985), metoda pravděpodobnostně vážených momentů (Hosking a Wallis, 1987) a regresní metoda založená na grafu výběrového průměru nad u proti hodnotě u (Davison a Smith, 1990). Diskuze vlastností těchto odhadů je např. i v Casstilo a Hadi (1997).

Smith (1985) uvádí, že metoda maximální věrohodnosti nejspíš nepovede k odhadu, pokud $|\xi| > 1$ a nemá obvyklé asymptotické vlastnosti pro $|\xi| > 0{,}5$. Pro $|\xi| < 0{,}5$ je již odhad asymptoticky normální.

Pro cenzorovaná data nejsou vlastnosti jednotlivých odhadů příliš prozkoumané. Článek Lin a Wang (2007) na základě simulační studie uvádí, že pro cenzorovaná data jsou maximálně věrohodné odhady spolehlivější než momentové odhady. Odhady v této práci jsou maximálně věrohodné.

3.9.3 Po částech spojená distribuční funkce

Při užití teorie extrémní hodnoty můžeme odhadnout pravý konec rozdělení zobecněným Paretovým rozdělením. Tohoto faktu můžeme využít ke zpřesnění odhadů konce rozdělení pomocí modelu po částech spojené distribuční funkce (v angličtině známý obvykle pod pojmem *spliced model*).

Model po částech spojené distribuční funkce je spojením odhadu distribuční funkce celého souboru a parametrického odhadu distribuční funkce jeho pravého konce. Distribuční funkce celého souboru může být odhadnuta neparametricky i parametricky. Tento model se rozvíjí v posledních letech zejména v pojišťovnictví a analýze rizik, viz třeba kapitola 3 v Dey a Jun (2016), kapitola 16 v Parodi (2014) nebo kapitola 1.5 v Peters a Shevchenko (2015). V České republice jej představil Vojtěch (2011).

Označíme distribuční funkci pro hlavní část (*bulk*) rozdělení $F(x|\boldsymbol{\theta}_\mathbf{b})$, kde $\boldsymbol{\theta}_\mathbf{b}$ označuje vektor parametrů modelu těla rozdělení, a distribuční funkci pravého konce $Z(x|\boldsymbol{\theta}_u)$, kde $\boldsymbol{\theta}_\mathbf{u}$ označuje vektor parametrů rozdělení pravého konce. Podíl pravého konce (tail fraction) označíme jako $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$, potom podle MacDonald a kol. (2011) můžeme psát distribuční funkci modelu po částech spojené distribuční funkce

$$F(x|\boldsymbol{\theta}_\mathbf{b}, \boldsymbol{\theta}_\mathbf{u}, \phi_u) = \begin{cases} \dfrac{1-\phi_u}{F(u|\boldsymbol{\theta}_\mathbf{b})} F(x|\boldsymbol{\theta}_\mathbf{b}) & \text{pro } x \le u \\ (1-\phi_u) + \phi_u Z(x|\boldsymbol{\theta}_\mathbf{u}) & \text{pro } x > u \end{cases} \quad (3.76)$$

Tuto směs nepřekrývajících se rozdělení označují Dey, Yan (2016) jako standardní model. Ve standardním modelu zlomek na prvním řádku normalizuje distribuční funkci tak, aby v prahu *u* odpovídala empirické distribuční funkci, a v tomto bodu navazuje odhad pravého konce. Pokud položíme $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$, používáme k odhadu distribuční funkce těla rozdělení přímo odhad tohoto rozdělení a odhad pravého konce navazuje z kvantilu tohoto rozdělení. V této práci volím druhý postup z toho důvodu, že v Turnbullově odhadu empirické distribuční funkce není možné najít odhad hodnoty kvantilu bez vyslovení určitého předpokladu o rozdělení pravděpodobnosti uvnitř odhadovaného intervalu (viz kapitola 3.3).

V kapitole 4.2 prezentuji nejprve odhad prahu pomocí následujícího optimalizačního postupu:

1. Hodnoty souboru jsem seřadil podle průměru z hodnot L_i a R_i.
2. Zvolil jsem práh *u* odpovídající 95% kvantilu parametrického odhadu na základě všech dat. Hodnota dolního nosiče je pevně stanovena rovnicí $\mu^* = u$.
3. Zbývající dva parametry zobecněného Paretova rozdělení jsem odhadl metodou maximální věrohodnosti.
4. Na základě tohoto odhadu jsem odhadnul střední hodnotu přesahu (*mean excess value*) jako střední hodnotu Paretova rozdělení:

$$E(X) = u^* + \frac{\sigma}{1-\xi}, \text{ pro } \xi < 1. \tag{3.77}$$

Pro hodnoty $\xi \geq 1$ není střední hodnota definovaná.

5. Zapsal jsem hodnoty odhadu ξ a střední hodnoty přesahu.
6. Hodnotu kvantilu použitého jako práh u jsem snížil o 1 p.b. a postup opakoval, dokud nezačala hodnota logaritmu věrohodnosti klesat a nejevila tendenci klesající trend změnit.
7. Pokud nebyl odhad parametru ξ významně odlišný od nuly, odhadoval jsem exponenciální rozdělení (viz rovnice (3.74)) od bodu, ve kterém byla maximální hodnota logaritmu věrohodnosti na obě strany, dokud nebyl rozdíl mezi oběma odhady statisticky významný na hladině významnosti 0,05 (pomocí testu věrohodnostním poměrem s jedním stupněm volnosti).
8. Jako konečný odhad jsem použil ten, při kterém byl logaritmus věrohodnosti nejvyšší. Pokud nebyl statisticky významný rozdíl mezi GPD a exponenciálním rozdělením, použil jsem exponenciální rozdělení.

Druhou alternativou tohoto postupu je přímá maximalizace věrohodnosti rozdělení popsaného rovnicí (3.76), kdy se souběžně odhadují parametry pravděpodobnostního rozdělení pro hlavní část, parametry zobecněného Paretovo rozdělení pro pravý konec a podíl konce $\phi_u = 1 - F(u|\boldsymbol{\theta}_\mathbf{b})$. Práh u potom odpovídá $100p\%$ kvantilu log-logistického rozdělení, kde $\phi_u = p$. Hodnota dolního nosiče je pevně stanovena rovnicí $\mu^* = u$.

V prvním postupu se tedy pro odhad hlavní části rozdělení používá i informace o konci tohoto rozdělení, ale samotný konec je odhadnut zvlášť. Ve druhém postupu se odhaduje hlavní část rozdělení pouze z empirických hodnot pod prahem u a konec rozdělení z empirických hodnot nad prahem rozdělení u.

V literatuře je popsán i třetí postup, který použiji pro srovnání. Nejedná se o po částech spojenou distribuční funkci, ale úvaha míří podobným směrem. Tímto postupem je plynulý přechod mezi modelem těla a modelem konce, tedy směs rozdělení s dynamickou váhou. V tomto modelu, v kontextu teorie extrémní hodnoty poprvé popsaném ve Frigessi a kol. (2003), se odhadují celá data jak pomocí vhodného rozdělení s tenkým koncem, tak zobecněným Paretovým rozdělením a oběma odhadům je přiřazena dynamická váha popsaná např. distribuční funkcí spojitého jednorozměrného rozdělení. Tato váha má zajistit, že hlavní část rozdělení je odhadováno především vybraným rozdělením a zobecněné Paretovo rozdělení

má na tento odhad malý vliv. U odhadu konce je to naopak. Hustota pravděpodobnosti takovéto směsi je popsána rovnicí:

$$v(t) = \frac{\big(1 - p(t;\boldsymbol{\theta})\big) f(t;\boldsymbol{\beta}) + p(t;\boldsymbol{\theta}) z\big(t;\xi,\sigma\big)}{Z(\boldsymbol{\theta},\boldsymbol{\beta},\xi,\sigma)}, \tag{3.78}$$

kde $f(t;\boldsymbol{\beta})$ je zvolená hustota pravděpodobnosti určená především pro nízké hodnoty, $z(t, \xi, \sigma)$ je hustota pravděpodobnosti zobecněného Paretova rozdělení, $p(t;\boldsymbol{\theta})$ je rostoucí dynamická váha nabývající hodnot (0, 1) a $Z(\boldsymbol{\theta}, \boldsymbol{\beta}, \xi, \sigma)$ je normalizační konstanta, která zajišťuje, aby integrál z hustoty pravděpodobnosti $v(t)$ byl roven 1.

V této práci používám pro popis rozdělení doby nezaměstnanosti log-logistické rozdělení, tedy $f(t, \boldsymbol{\beta})$ bude hustotou pravděpodobnosti log-logistického rozdělení. Jako dynamickou váhu volím ve shodě s Frigessi a kol. (2003) distribuční funkci Cauchyho rozdělení, tedy

$$p(t,\boldsymbol{\theta}) = \frac{1}{2} + \frac{1}{\pi} arctg\left(\frac{t-\mu}{\sigma}\right), \qquad \boldsymbol{\theta} = (\mu,\sigma), \quad \mu,\sigma > 0. \tag{3.79}$$

Tato váha zajišťuje základní vlastnost, tedy je rostoucí a v intervalu (0, 1). Parametr μ je zároveň mediánem Cauchyho rozdělení, tedy se jedná o hodnotu, kdy je váha rozdělená půl na půl mezi log-logistické a zobecněné Paretovo rozdělení.

Pokud je jako váha zvolena Heavisideova funkce, odpovídá tento dynamický model standardnímu modelu v rovnici (3.72), viz Scarrott a MacDonald (2012).

3.10 Software

Analýza zprava cenzorovaných dat je obecně přístupná v mnoha běžně dostupných statistických softwarech, namátkou v SPSS, SAS, Stata či S-PLUS. Intervalově cenzorovaná data však obvykle nejsou implementována popřípadě jen částečně nebo s řadou omezení, např. v SAS, viz So a kol. (2010).

V práci proto kromě Excelu přednostně používám software *R* (R Core Team, 2016) s následujícími balíčky[6]:

[6] Veškeré balíčky týkající se analýzy přežití jsou k nalezení na adrese https://cran.r-project.org/web/views/Survival.html

- *survival* (Therneau, 2015 a Therneau a Grambsch, 2000) je obecně nejuniverzálnějším balíčkem pro analýzu cenzorovaných dat, který umí pracovat i s intervalově cenzorovanými daty. Mezi implementované metody obecně patří AFT model, Coxův model a neparametrické modely včetně Turnbullova odhadu. Coxův model ovšem není určen pro intervalově cenzorovaná data,
- *interval* (Fay a Shaw, 2011) je balíček zaměřený na Turnbullův odhad a neparametrické testy pro intervalově cenzorovaná data včetně váženého log-rank testu,
- *fitdistrplus* (Muller a kol., 2015) je balíček aplikující metodu maximální věrohodnosti na odhad libovolného pravděpodobnostního rozdělení včetně těch definovaných uživatelem. Odhad je aplikovatelný i na intervalově cenzorovaná data. Vedle asymptotického odhadu intervalu spolehlivosti umožňuje i aplikaci metody bootstrap,
- *fExtremes* (Wuertz a kol., 2013) je balíček určený primárně pro aplikaci teorie extrémních hodnot na necenzorovaných datech. Z tohoto balíčku jsem použil definici zobecněného Paretova rozdělení pro užití v balíčku *Fitdistrplus*.
- *coefplot* (Lander, 2016) a *arm* (Gelman a Su, 2016) jsou balíčky sloužící k vykreslení intervalového odhadu parametrů z regresních modelů, popřípadě vlastnoručně zadaných údajů.

4 Výsledky

V této kapitole představuji výsledky jednotlivých modelů. Nejprve v podkapitole 4.1 je představeno celkové rozdělení doby nezaměstnanosti v jednotlivých obdobích a testována odlišnost. V podkapitole 4.2 se zabývám modelováním pravého konce prvního datového souboru. V podkapitole 4.3 se zabývám modelováním vztahu jednotlivých proměnných k době do nalezení práce či době nezaměstnanost. V kapitole 4.4 potom představuji úplný a minimální adekvátní model doby nezaměstnanosti a v kapitole 4.5 se zabývám interakcemi proměnných ve vztahu k době do nalezení práce.

Jedním z důležitých kroků při budování regresního AFT modelu je rozhodnutí o užitém pravděpodobnostním rozdělení. Toto rozhodnutí jsem zde učinil na základě věrohodnosti, které má dané rozdělení v modelu rozdělení doby do nalezení zaměstnání či doby nezaměstnanosti pro celý soubor dělený podle období. Hodnoty věrohodnosti takového modelu pro jednotlivá zvažovaná parametrická rozdělení jsou v tabulce 4.1.

Tabulka 4.1 Vybraná rozdělení a logaritmus věrohodnosti AFT modelu podle období

Rozdělení	soubor 1	soubor 2	soubor 3
Log-logistické	-3 974,0	-8 277,5	-9 446,9
Log-normální	-3 976,4	-8 362,9	-9 465,0
Weibullovo	-4 196,7	-8 434,9	-9 454,5
Exponenciální	-4 201,3	-8 487,7	-9 577,4
Studentovo	-4 769,8	-10 657,7	-11 480

Z vybraných rozdělení mají log-logistické a log-normální nejvíce flexibilní rizikovou funkci, která může dle hodnoty parametrů být monotónní klesající nebo kopcovitá. Weibullovo rozdělení má oproti tomu jen monotónní rizikovou funkci a exponenciální rozdělení má riziko konstantní. Pokud předpokládáme, že riziko nalezení zaměstnání má nejdříve rostoucí a poté klesající tendenci, jeví se volba log-normálního nebo log-logistického rozdělení logická.

Vymykající se je výsledek u souboru 3, ve kterém má model s Weibullovým rozdělením vyšší věrohodnost než model s log-normálním rozdělením. Zdá se, že v souboru 3 se možná potkáváme s klesající rizikovou funkcí, tedy že riziko nalezení zaměstnání s rostoucí dobou nezaměstnanosti klesá. Připomínám zde, že v souboru 3 nejde o dobu do nalezení zaměstnání.

Tuto úvahu podporují i výsledky v tabulkách 4.10 a 4.11, ve kterých se modus doby nezaměstnanosti pohybuje kolem jednoho měsíce, což je nižší hodnota než v souborech 1 a 2.

4.1 Celkové rozdělení

V souboru 1 byla nejkratší doba hledání práce před krizí, zatímco v době krize a po krizi se doba hledání práce neliší, je delší. Liší se zde závěry o střední době nezaměstnanosti z neparametrického a parametrického odhadu (viz tabulky 4.2 a 4.4), což je dáno ukončením neparametrického odhadu. Pokud zkonstruujeme neparametrický odhad a pro odhad střední hodnoty jej ukončíme dřív, vychází výsledky v souladu s AFT modelem – viz tabulka 4.3, kde byl střední bod posledního intervalu pevně stanoven na hodnotu 20. Důvod je zřetelně vidět i v obrázku 4.1, kde hodnoty funkce přežití pro předkrizové období jsou pro většinu hodnot *t* nižší než hodnoty krizové a pokrizové, což se obrací na konci rozdělení. Doba do nalezení práce se během krize a po krizi prodloužila o cca 18 % oproti předkrizovému období.

Tabulka 4.2 Turnbullův odhad a log-rank test pro soubor 1

Období	Rmean	s.e
Před krizí	9,9	0,520
Během krize	9,94	0,317
Po krizi	10,8	0,374
ChiSq	18,43	
Df	2	
p-hodnota	0,0001	

Tabulka 4.3 Turnbullův odhad (omezení = 20)

Období	rmean	s.e.
Před krizí	7,58	0,254
Během krize	8,49	0,180
Po krizi	8,59	0,189

V tabulce 4.4 je specifikován AFT model pro celý soubor 1. Tento formát výstupu se bude vyskytovat i nadále, proto jej podrobněji popíšu zde. Ve sloupci *Odhad* se vyskytuje odhad

příslušného parametru, ve sloupci *z* je hodnota statistiky Waldova testu (rovnice (3.51)) a v sloupci p-hodnota je p-hodnota Waldova testu. Sloupce $t_{0,5}$ a $t_{0,9}$ obsahují 50% a 90% kvantil doby do nalezení práce. Sloupec *E*(*T*) popisuje odhadnutou střední hodnotu a *Modus* pak odhadnutý modus. Sloupce *s.e.* obsahují směrodatné chyby odhadu pro odhady ve sloupcích vlevo od nich.

V tabulce 4.4 je tedy odhadnutý medián doby do nalezení práce před krizí 5,58 měsíců, během krize a po krizi 6,61 měsíců. 90% kvantil je odhadnut na 17,23 měsíců respektive 20,38 měsíců. Odhady střední hodnoty doby do nalezení práce jsou 9 měsíců před krizí a 10,65 měsíců po krizi. Modus jsem odhadnul na 3,12 měsíců před krizí a 3,70 měsíců během a po krizi.

Tabulka 4.4 AFT model dle období pro soubor 1

Období	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ – Před krizí	1,720	0,0402	42,79	< 0,0001	5,58	0,22	17,23	0,77	9,00	3,12
Během krize	0,168	0,0492	3,41	0,0006	6,61	0,19	20,38	0,70	10,65	3,70
Po krizi	0,168	0,0496	3,39	0,0007	6,61	0,19	20,38	0,72	10,65	3,70
ln(σ)	– 0,668	0,0184	– 36,30	< 0,0001						

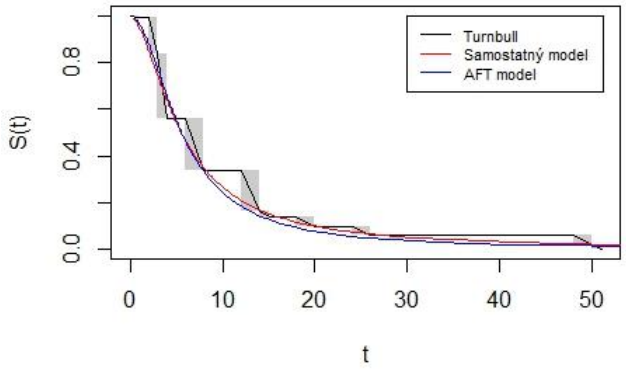
Tabulka 4.5 obsahuje odhady pro rozdělení doby do nalezení práce v souboru 1 pro každé období zvlášť. Můžeme srovnat tyto odhady s odhady v tabulce 4.4. Odhady mediánů se liší v setinách měsíce, vyšší odchylky jsou pro odhady 90% kvantilů, kdy především pro období před krizí se odhad liší o 2,5 měsíce. Modus je naopak před krizí odhadnut dřív o 0,5 měsíce. Odchylky pro další dvě období nejsou již tolik výrazné.

Grafické srovnání těchto dvou odhadů s Turnbullovým odhadem pro jednotlivá období nabízí postupně obrázky 4.1 až 4.3. Na nich je vidět podobná informace, tedy že nejvíce se samostatný parametrický odhad od AFT modelu odchyluje odhad pro období před krizí.

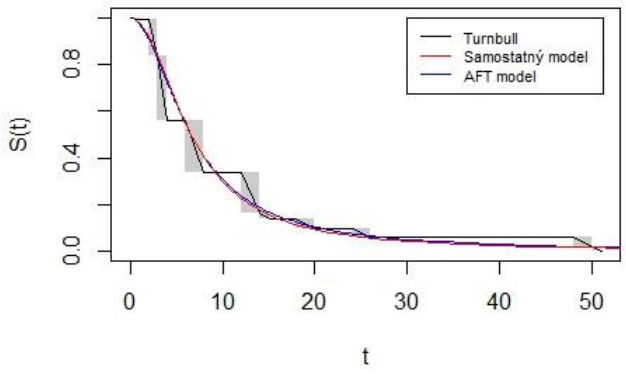
Na těchto obrázcích je Turnbullův odhad znázorněn černou linií s šedými obdélníky. Tyto obdélníky vyjadřují nejednoznačnost Turnbullova odhadu uvnitř jednotlivých intervalu, viz diskuze v kapitole 3.3.

Tabulka 4.5 Parametrické odhady log-logistického rozdělní pro tři období, soubor 1

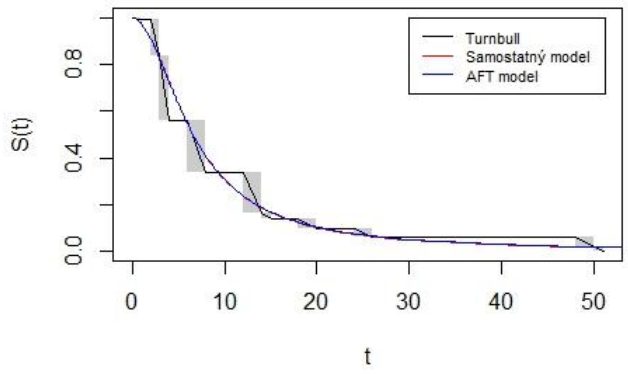
Před krizí	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	1,714	0,0440	38,95	< 0,0001	5,55	0,24	19,77	1,32	10,39	2,59
ln(σ)	– 0,548	0,0410	– 13,37	< 0,0001						
Během krize	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	1,890	0,0440	42,95	< 0,0001	6,62	0,18	19,23	0,77	10,10	3,96
ln(σ)	-0,723	0,0410	-17,63	< 0,0001						
Po krizi	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	1,888	0,0440	42,91	< 0,0001	6,61	0,19	20,13	0,86	10,53	3,75
ln(σ)	-0,679	0,0410	-16,56	< 0,0001						



Obrázek 4.1 Srovnání odhadů pro období před krizí, soubor 1

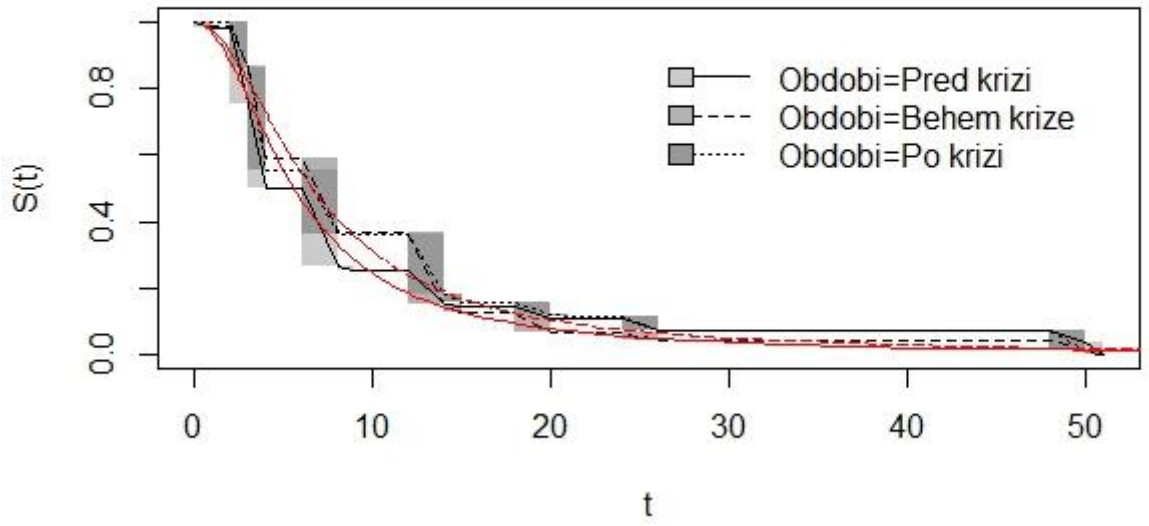


Obrázek 4.2 Srovnání odhadů pro období během krize, soubor 1



Obrázek 4.3 Srovnání odhadů pro období po krizi, soubor 1

Obrázek 4.4 nabízí pohled na rozdíl mezi jednotlivými obdobími v Turnbullově odhadu a AFT modelu. Je zde vidět podobnost v pravděpodobnostním rozdělení doby do nalezení práce během krize a po krizi a odlišnost tohoto rozdělení v období před krizí, kdy byla funkce přežití nižší, tedy doba do nalezení práce byla kratší.



Obrázek 4.4 Turnbullův odhad a AFT model pro nalezenou práci

V souboru 2 byla doba do nalezení práce naopak nejvyšší v předkrizovém období, což je vidět v neparametrickém i parametrickém odhadu v tabulkách 4.6 a 4.7. Opět se nijak

významně neliší celopopulační rozdělení v období krize a po krizi. Doba do nalezení práce byla tedy odhadnuta kratší o zhruba 21 %.

V tabulkách 4.7 a 4.8 můžeme opět porovnat odhady AFT modelu a parametrické odhady aplikované na samostatné soubory v jednotlivých obdobích.

Tabulka 4.6 Turnbullův odhad a log-rank test pro soubor 1

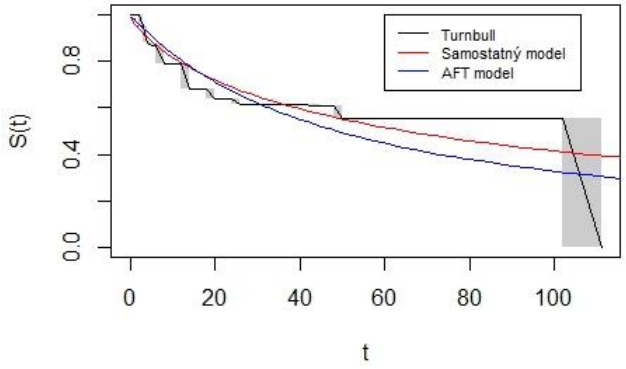
Období	Rmean	s.e
Před krizí	70,3	0,520
Během krize	66,5	0,317
Po krizi	66,3	0,374
ChiSq	24,50	
Df	2	
p-hodnota	<0,0001	

Tabulka 4.7 AFT model pro soubor 2

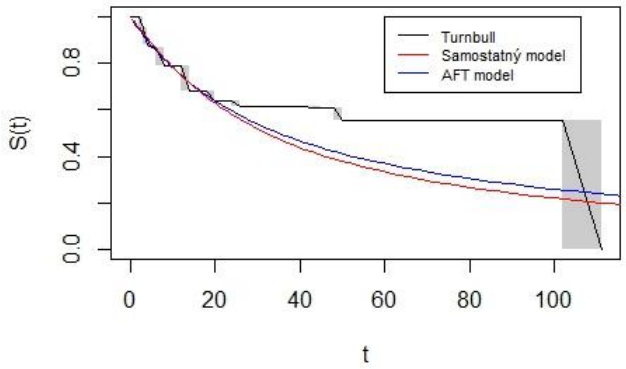
Období	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ – Před krizí	3,857	0,0526	73,33	< 0,0001	47,32	2,61	417,49	31,9	200,37	11,18
Během krize	–0,238	0,0604	– 3,94	0,0001	37,30	1,43	329,07	20,2	157,94	8,81
Po krizi	–0,228	0,0610	– 3,74	0,0002	37,68	1,49	332,38	20,6	159,52	8,90
ln(σ)	0,530	0,0162	32,72	< 0,0001						

Tabulka 4.8 Parametrické odhady log-logistického rozdělní pro tři období, soubor 2

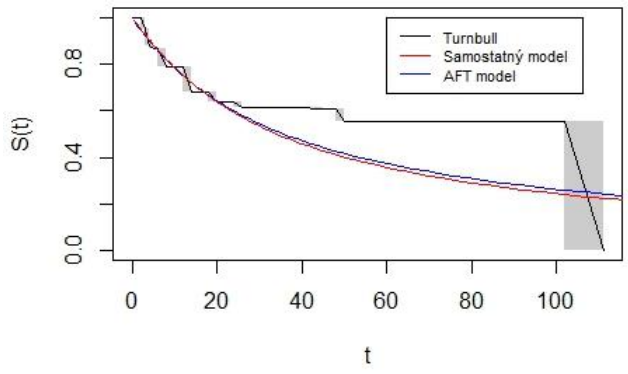
Před krizí	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	4,173	0,0818	51,01	< 0,0001	64,91	5,31	1025,47	173,28	NA	NA
ln(σ)	0,228	0,0378	6,03	< 0,0001						
Během krize	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	3,466	0,0417	83,12	< 0,0001	32,01	1,33	230,50	19,38	288,28	2,30
ln(σ)	–0,107	0,0268	– 3,99	0,0001						
Po krizi	Odhad	s.e.	z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	3,533	0,0437	80,76	<0,0001	34,23	1,47	226,39	18,16	216,92	3,71
ln(σ)	–0,058	0,0266	–2,17	0,0297						



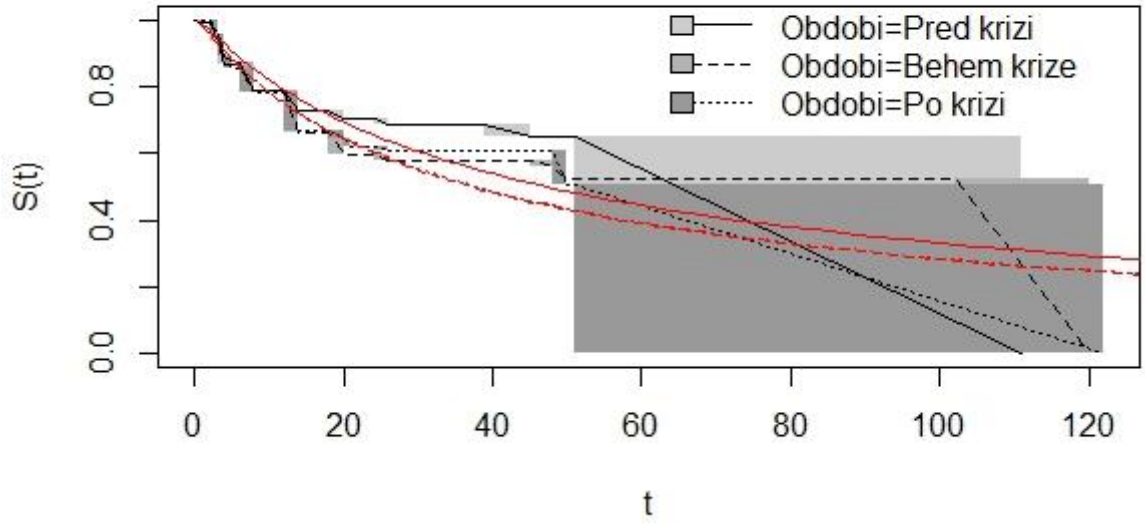
Obrázek 4.5 Srovnání odhadů pro období před krizí, soubor 2



Obrázek 4.6 Srovnání odhadů pro období během krize, soubor 2



Obrázek 4.7 Srovnání odhadů pro období po krizi, soubor 2



Obrázek 4.8 Turnbullův odhad a AFT model pro soubor 2

V souboru 3 vychází odlišně období krize od období předkrizového i pokrizového, které se naopak statisticky významně neliší. Jak je vidět v tabulce 4.8, tak v období krize opět z hlediska doby nezaměstnanosti vychází, že lidé práci hledají v průměru o cca 14 % kratší dobu než před krizí. Doba hledání práce po krizi je kratší o cca 8 %, ale jak bylo napsáno, rozdíl není statisticky významný. Obdobný vztah s neparametrickým odhadem střední hodnoty lze vidět i v tabulce 4.7.

Tabulka 4.9 Turnbullův odhad a log-rank test pro soubor 3

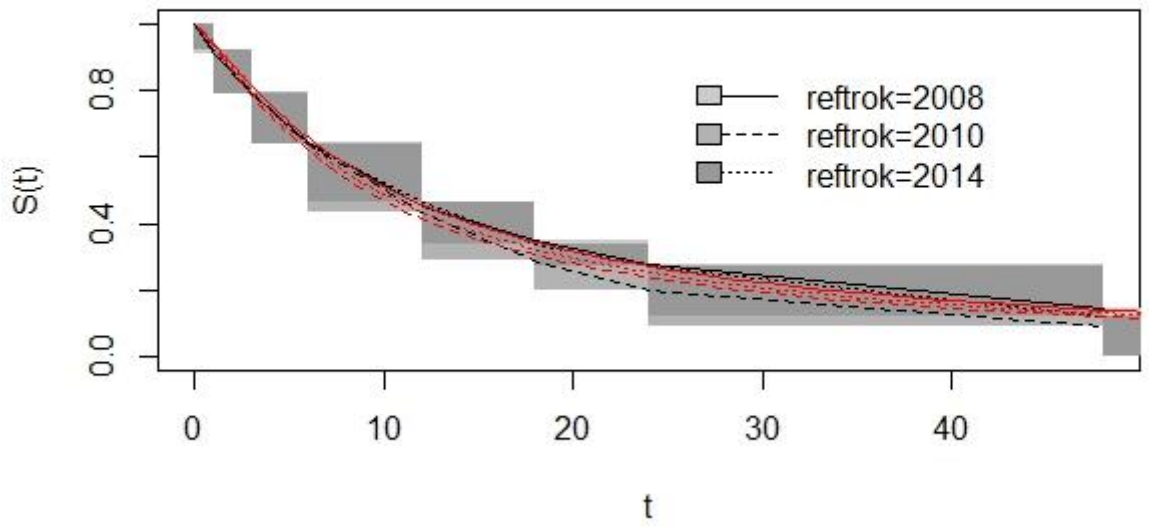
Období	Rmean	s.e
2008	17,8	0,482
2010	15,2	0,335
2014	16,7	0,398
ChiSq	17,06	
Df	2	
p-hodnota	0,0002	

Tabulka 4.10 AFT model pro soubor 3

Období	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ – 2008	2,341	0,0436	53,69	< 0,0001	10,39	0,45	66,21	3,34	58,05	1,31
2010	– 0,147	0,0550	– 2,67	0,0075	8,97	0,30	57,16	2,38	50,12	1,13
2014	– 0,082	0,0575	– 1,43	0,1538	9,57	0,36	61,00	2,75	53,48	1,20
ln(σ)	– 0,171	0,0137	– 12,48	< 0,0001						

Srovnání Turnbullova odhadu a AFT modelu pro všechna tři období dohromady je v obrázku 4.9. Tento obrázek napovídá, že rozdíl v AFT modelu je oproti neparametrickému odhadu nižší.

Srovnání odhadů v tabulkách 4.10 (AFT model) a 4.11 (jednotlivé parametrické odhady) opět ukazuje velmi podobný medián a výraznější odchylky v odhadech 90% kvantilu a střední hodnoty, zejména pro období před krizí, ve kterém byla střední hodnota odhadnutá samostatným odhadem přes 110 měsíců oproti 58 měsícům v AFT modelu. Odhad mediánu se zde lišil o skoro 11 měsíců, taktéž největší rozdíl za všechna tři období.

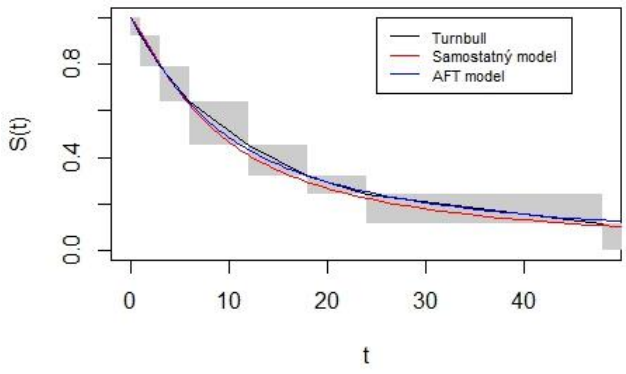


Obrázek 4.9 Turnbullův odhad a AFT model pro soubor 3

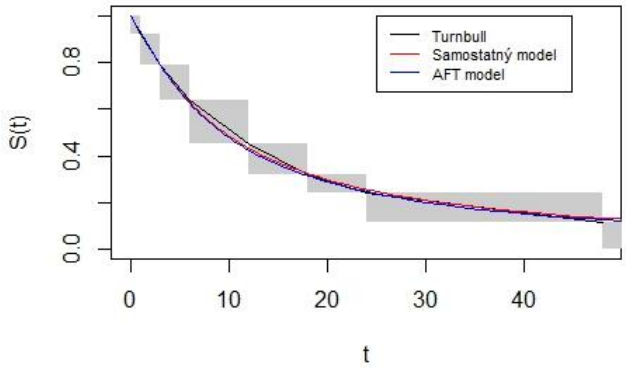
Srovnání Turnbullova odhadu, AFT modelu a parametrického odhadu pro jednotlivá čtvrtletí jsou v obrázcích 4.10 až 4.12

Tabulka 4.11 Parametrické odhady log-logistického rozdělní pro tři období, soubor 3

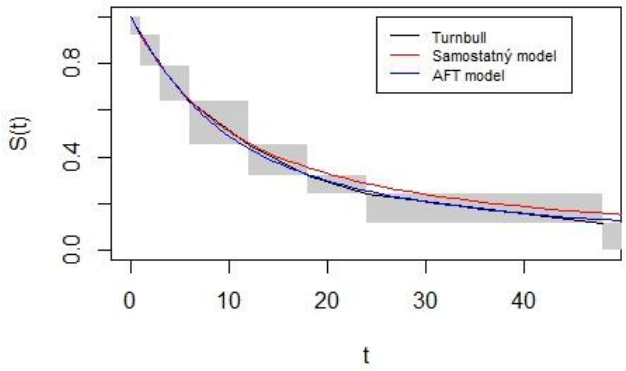
2008	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	2,341	0,0464	50,45	< 0,0001	10,39	0,48	77,25	5,60	110,45	0,62
ln(σ)	– 0,091	0,0278	– 3,27	0,0011						
2010	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	2,198	0,0319	68,90	< 0,0001	9,01	0,29	50,72	2,43	35,83	1,69
ln(σ)	– 0,240	0,0215	– 11,16	< 0,0001						
Po krizi	Odhad	s.e.	Z	p-hodnota	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
μ	2,259	0,0380	59,45	< 0,0001	9,57	0,36	63,09	3,64	59,87	1,05
ln(σ)	– 0,153	0,0235	– 6,51	< 0,0001						



Obrázek 4.10 Srovnání odhadů pro období před krizí, soubor 3



Obrázek 4.11 Srovnání odhadů pro období během krize, soubor 3

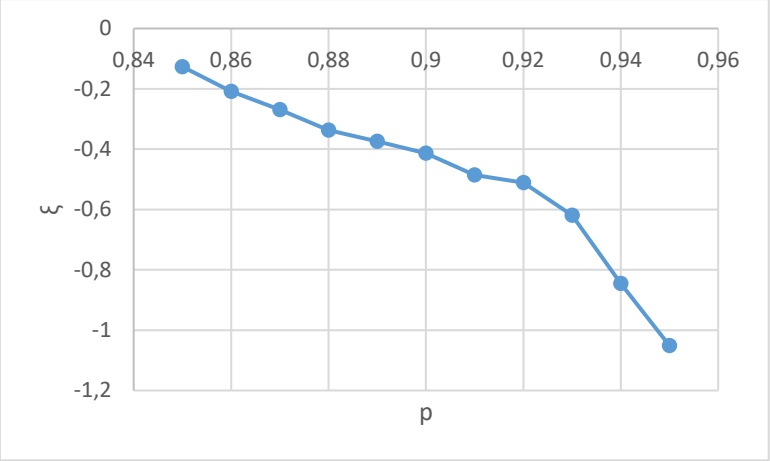


Obrázek 4.12 Srovnání odhadů pro období po krizi, soubor 3

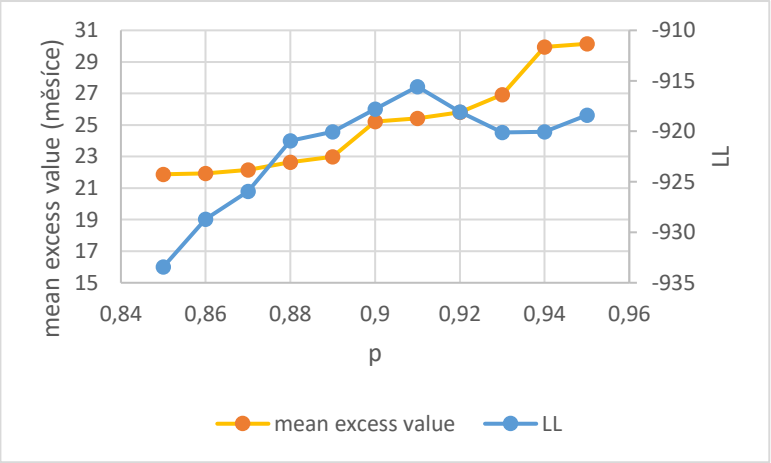
Při zkoumání jednotlivých souborů jsem zaznamenal odlišné výsledky, když v souboru 1 byla doba do nalezení práce během krize a po krizi delší než před ní, zatímco v souborech 2 a 3 to bylo obráceně. Zatímco v souboru 1 je zaznamenána doba do nalezení práce pouze u těch, kteří ji skutečně nalezli, v dalších souborech se jedná o dobu do nalezení zaměstnání všech a dobu nezaměstnanosti. Zde předpokládám efekt nových nezaměstnaných, kdy nástup krize zvýšil míru nezaměstnanosti a noví nezaměstnaní pak snížili střední dobu nezaměstnanosti celkově. Naopak prodloužená doba do nalezení práce v souboru 1 odráží těžší pozici uchazečů o práci, kteří ji skutečně hledali déle.

4.2 Rozdělení konce

V této kapitole představuji odhad pravého konce rozdělení doby nezaměstnanosti ze souboru 1 popsaný obecně v kapitole 3.9.3. Pro každé období provádím odhad celkem 3 modelů a tyto mezi sebou porovnávám Akaikeho informačním kritériem a Bayesovským informačním kritériem. Prvním modelem je log-logistický odhad celého rozdělení a nahrazení jeho konce zobecněným Paretovým rozdělením od určitého prahu, který je určen pomocí věrohodnosti (LL+GPD). Druhým modelem je souběžný odhad log-logistického a zobecněného Paretova rozdělení s prahem (LL+GPD+μ*). Třetím modelem je směs log-logistického a Paretova rozdělení s dynamickou váhou, kterou představuje distribuční funkce Cauchyho rozdělení (LL+GPD+C).



Obrázek 4.13 Odhad parametru ξ pro různé hodnoty *p*, před krizí



Obrázek 4.14 Odhad střední hodnoty přesahu (mean excess value) a hodnota logaritmu věrohodnostní funkce (LL), pro různá *p*, před krizí

V období před krizí při hledání prahu zobecněného Paretova rozdělení je odhad parametru ξ klesající pro rostoucí procento použitého kvantilu, od 92% kvantilu pak prudce klesající, viz obrázek 4.13. Podobný vývoj opačným směrem ukazuje střední hodnota přesahu, z čehož lze usuzovat na to, že po tento kvantil je vhodné určit práh, viz obrázek 4.14. Věrohodnost modelu je nejvyšší v 91% kvantilu, tedy jako práh jsem zde použil 91% kvantil log-logistického rozdělení, jehož hodnotou je 20,972 měsíců.

Jelikož ve druhém modelu není odhad parametru ξ statisticky významný, provedl jsem odhad pravého konce pomocí exponenciálního rozdělení, tento model značím LL+EXP+μ*.

Ve třetím modelu je opět statisticky nevýznamný odhad parametru ξ, tabulka tedy poskytuje odhad s exponenciálním rozdělením namísto zobecněného Paretova. Tento model značím jako LL+EXP+C.

Tabulka 4.12 Srovnání odhadů, před krizí

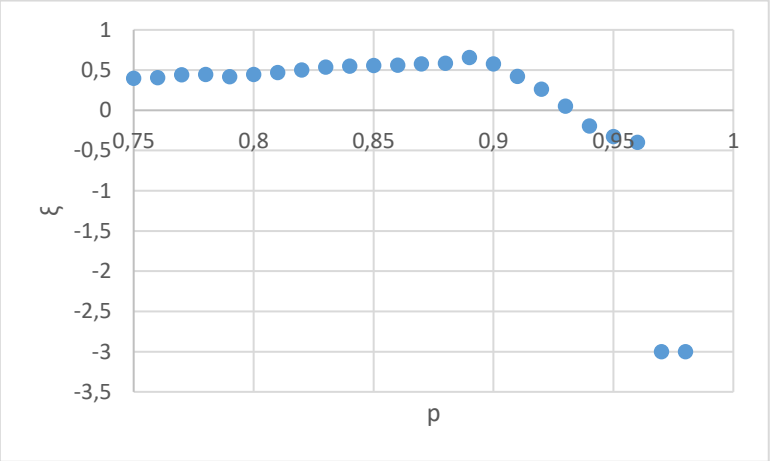
Model	LL	Parametrů	AIC	BIC
LL+ EXP + μ*	-874,29	4	1756,58	1774,325
LL + GPD + μ*	-873,87	5	1757,74	1779,921
LL + GPD	-915,57	4	1839,14	1856,885
LL + EXP + C	-915,24	5	1840,48	1862,661
LL + GPD + C	-922,47	6	1856,94	1883,557
LL	-935,06	2	1874,12	1882,992

Tři uvedené odhadnuté modely jsou uveden v tabulce 4.13, srovnání všech šesti modelů je v tabulce 4.12. V této tabulce jsou odhady seřazeny podle hodnoty Akaikeho informačního kritéria AIC. Za nejvhodnější model lze označit model LL+EXP+μ*, ve kterém se simultánně odhadovala jak hlavní část, tak konec rozdělení i práh, přičemž konec rozdělení se modeloval exponenciálním rozdělením.

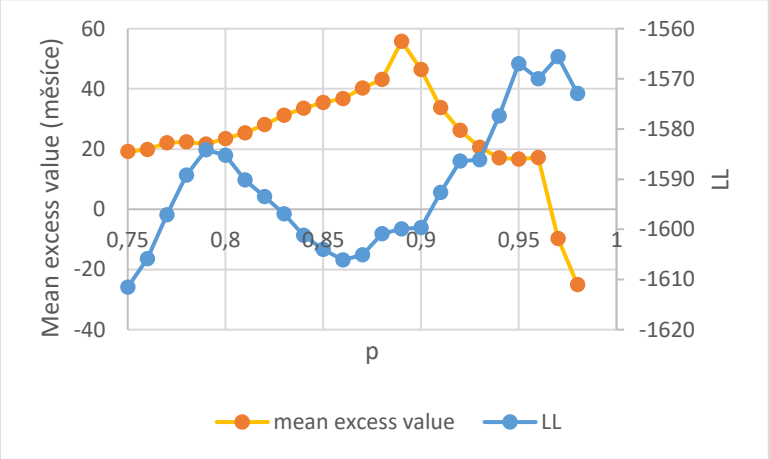
Tabulka 4.13 Odhad LL+GPD pro období před krizí

LL+GPD	Odhad	s.e.	Z	p-hodnota
α	5,460	0,8402	6,50	< 0,0001
β	1,709	0,0704	24,28	< 0,0001
ξ	– 0,486	0,1750	– 2,78	0,0027
μ*	21,150	NA	NA	NA
σ	48,058	7,5460	6,37	< 0,0001
LL + EXP + μ*	Odhad	s.e.	Z	p-hodnota
α	4,393	0,7543	5,82	< 0,0001
β	2,817	0,0904	31,16	< 0,0001
ξ	0	NA	NA	NA
μ*	7,568	1,1353	6,67	< 0,0001
σ	30,396	8,3430	3,64	0,0001
LL + EXP + C	Odhad	s.e.	Z	p-hodnota
α	4,840	0,7431	6,51	< 0,0001
β	2,368	0,1043	22,70	< 0,0001
ξ	0	NA	NA	NA
μ*	0	NA	NA	NA
σ	19,997	4,3324	4,62	< 0,0001
μ	56,933	7,4324	7,66	< 0,0001
τ	5,804	1,0954	5,30	< 0,0001

V období během krize při hledání prahu zobecněného Paretova rozdělení je odhad parametru ξ poměrně konstantní pro rostoucí procento použitého kvantilu, od 90% kvantilu je pak klesající, viz obrázek 4.15. Střední hodnota přesahu je rostoucí po 89% kvantil, následuje její pokles po 96% kvantil a poté prudký pád, viz obrázek 4.16. Věrohodnost modelu je nejvyšší v 97% kvantilu, nicméně odhad parametru ξ je zde nižší než – 3, tedy jako práh jsem zde použil 95% kvantil log-logistického rozdělení (druhá nejvyšší věrohodnost), jehož hodnota je 26,934 měsíců.



Obrázek 4.15 Odhad parametru ξ pro různé hodnoty *p*, během krize



Obrázek 4.16 Odhad střední hodnoty přesahu (mean excess value) a hodnota logaritmu věrohodnostní funkce (LL), pro různá *p*, před krizí

Tabulka 4.14 Odhad LL+GPD pro období během krize

LL + GPD	Odhad	s.e.	Z	p-hodnota
α;	6,677	1,1432	5,84	< 0,0001
β	2,111	0,0893	23,64	< 0,0001
ξ	– 0,326	0,1534	– 2,13	0,0168
μ*	26,934	NA	NA	NA
σ	30,887	7,4230	4,16	< 0,0001
LL + GPD + μ*	Odhad	s.e.	Z	p-hodnota
α	6,267	0,9423	6,65	< 0,0001
β	2,394	0,1012	23,66	< 0,0001
ξ	– 0,251	0,1423	– 1,76	0,0389
μ*	21,408	1,4323	14,95	< 0,0001
σ	33,174	5,4324	6,11	< 0,0001
LL + EXP + C	Odhad	s.e.	Z	p-hodnota
α	6,508	1,1234	5,79	< 0,0001
β	2,317	0,0932	24,86	< 0,0001
ξ	0	NA	NA	NA
μ*	0	NA	NA	NA
σ	17,314	3,2345	5,35	< 0,0001
μ	77,908	5,4256	14,36	< 0,0001
τ	4,024	1,2345	3,26	< 0,0001

Odhady všech modelů jsou popsány v tabulce 4.14. Ve třetím modelu je opět statisticky nevýznamný odhad parametru ξ, tato tabulka tedy poskytuje odhad s exponenciálním rozdělením namísto zobecněného Paretova.

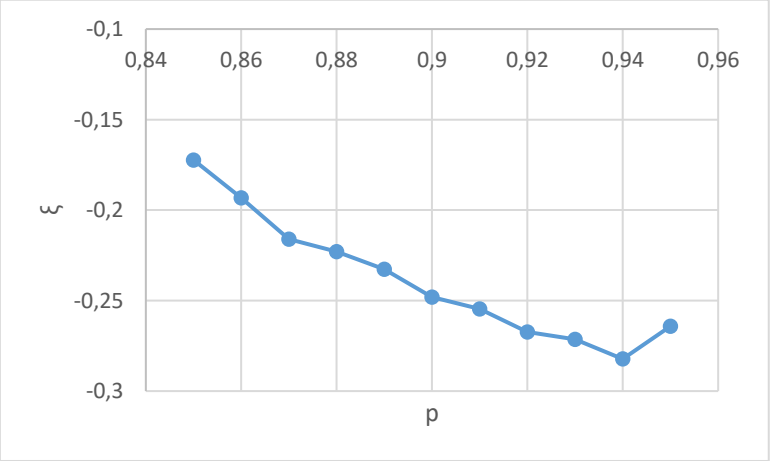
Srovnání všech pěti odhadů je v tabulce 4.15 V této tabulce jsou odhady seřazeny podle hodnoty Akaikeho informačního kritéria AIC. Za nejvhodnější model lze označit model LL+GPD+μ*, ve kterém se simultánně odhadovala jak hlavní část, tak konec rozdělení i práh, přičemž konec rozdělení se modeloval zobecněným Paretovým rozdělením. Na druhou stranu posuzování podle Bayesovského informačního kritéria by vybralo model s exponenciálním

koncem a jako druhý nejvhodnější by označilo čistě log-logistický model. V období během krize samostatný odhad konce rozdělení vylepšuje celkový odhad nepatrně.

Tabulka 4.15 Srovnání odhadů, během krize

Model	LL	parametrů	AIC	BIC
LL + GPD + μ*	-1558,00	5	3126,00	3150,872
LL + EXP + μ*	-1560,45	4	3128,9	3148,798
LL	-1567,57	2	3139,14	3149,089
LL + GPD	-1566,98	4	3141,96	3161,858
LL + EXP + C	-1567,25	5	3144,50	3169,372
LL + GPD + C	-1606,04	6	3224,08	3253,927

V období po krizi při hledání prahu zobecněného Paretova rozdělení je odhad parametru ξ klesající pro rostoucí procento použitého kvantilu až do 94% kvantilu viz obrázek 4.17. Střední hodnota přesahu je rostoucí do 91% kvantilu a následně klesající, což znázorňuje obrázek 4.18. Věrohodnost modelu je nejvyšší taktéž v 91% kvantilu, tedy jako práh jsem zde použil 91% kvantil log-logistického rozdělení, jehož hodnotou je 21,522 měsíců.



Obrázek 4.17 Odhad parametru ξ pro různé hodnoty *p*, po krizi

Tabulka 4.16 Odhad LL+GPD pro období po krizi

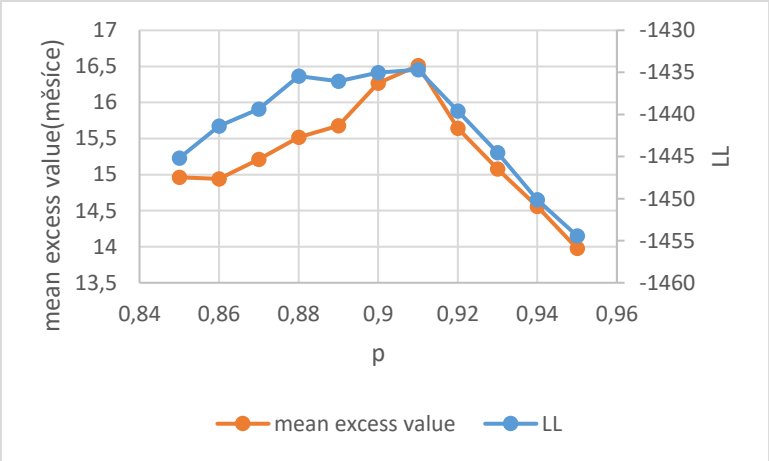
LL + GPD	Odhad	s.e.	Z	p-hodnota
α	6,567	0,9343	7,03	< 0,0001
β	1,949	0,0783	24,89	< 0,0001
ξ	– 0,255	0,1222	– 2,09	0,0185
μ*	21,522	NA	NA	NA
σ	26,195	6,3242	4,14	< 0,0001
LL + EXP + μ*	Odhad	s.e.	Z	p-hodnota
α	3,414	0,8535	4,00	< 0,0001
β	9,766	0,5643	17,31	< 0,0001
ξ	0	NA	NA	NA
μ*	3,436	1,2324	2,79	0,0027
σ	14,766	3,2324	4,57	< 0,0001
LL + EXP + C	Odhad	s.e.	Z	p-hodnota
α	5,627	0,9423	5,97	< 0,0001
β	2,281	0,1205	18,93	< 0,0001
ξ	0	NA	NA	NA
μ*	0	NA	NA	NA
σ	16,655	4,4256	3,76	0,0001
μ	54,169	7,0965	7,63	< 0,0001
γ	4,991	1,6457	3,03	0,0012

Odhady tří vybraných modelů (při nevýznamnosti parametru ξ jsem opět použil exponenciální rozdělení) jsou v tabulce 4.16.

Srovnání všech šesti modelů je v tabulce 4.17. V této tabulce jsou odhady seřazeny podle hodnoty Akaikeho informačního kritéria *AIC*. Za nejvhodnější model lze označit model LL+EXP+μ* stejně jako v období před krizí.

Tabulka 4.17 Srovnání odhadů, před krizí

Parametr	LL	parametrů	AIC	BIC
LL + EXP + μ*	-1196,36	4	2400,72	2420,204
LL + GPD + μ*	-1248,53	5	2507,06	2531,415
LL + GPD	-1434,68	4	2877,36	2896,844
LL + EXP + C	-1437,11	5	2884,22	2908,575
LL	-1449,37	2	2902,74	2912,482
LL + GPD + C	-1479,31	6	2970,62	2999,847



Obrázek 4.18 Odhad střední hodnoty přesahu (mean excess value) a hodnota logaritmu věrohodnostní funkce (LL), pro různá *p*, po krizi

Z výše uvedených odhadů je vidět, že v období před krizí a po krizi lze vylepšit odhad pravých konců doby nezaměstnanosti pomocí teorie extrémní odhady, nejlépe přímým odhadem modelu z rovnice (3.75). Pravé konce tíhnou k exponenciálnímu rozdělení, které se uplatňuje před krizí od zhruba 7,5 měsíců, zatímco po krizi již od necelých 3,5 měsíců. V období během krize je odhad pravého konce vylepšen nepatrně, přičemž zobecněné Paretovo rozdělení se zde uplatňuje až po dvacátém měsíci.

Použití směsi se neukázalo na těchto datech jako příliš vhodné, rozdělení chvostu převládá až od 57., 78. respektive 54. měsíce. Tyto údaje jsou nicméně kompatibilní s poznatkem, že pravý konec nebyl během krize log-logistickým rozdělením podceněn.

4.3 Jednotlivé proměnné

Při zkoumání změn podle jednotlivých proměnných sledují vývoj vztahu doby nezaměstnanosti k těmto proměnným jednotlivě pomocí log-rank testů a AFT modelů. Zkoumám i vliv období na jednotlivé hodnoty těchto proměnných, tedy zda se statisticky významně liší hodnoty jednotlivých proměnných v různých obdobích, a to pomocí log-rank testů. Zvolená hladina významnosti činí 5 %.

S ohledem na množství vygenerovaných výstupů jsou kompletní modely k dispozici v elektronické příloze A, v příloze B pak reprodukuji AFT modely transformací na akcelerační faktor $\exp(x^T b)$. 95% interval spolehlivosti můžu transformovat

$$\begin{aligned} &95\%\ \text{I.S.} = \left(b - s.e.(b) * u_{0,975} < \beta < b + s.e.(b) * u_{0,975}\right) \\ &95\%\ \text{I.S.} = \left(\exp(b) / \exp\left(s.e.(b) * u_{0,975}\right) < \exp(\beta) < \exp(b) * \exp\left(s.e.(b) * u_{0,975}\right)\right), \end{aligned} \tag{4.1}$$

kde b je bodový odhad parametru β, s.e.(b) je směrodatná chyba odhadu a $u_{0,975}$ je 97,5% kvantil normovaného normálního rozdělení. Výraz $\exp(s.e.(b)*u_{0,975})$ potom slouží k určení 95% intervalu spolehlivosti pro odhad akceleračního faktoru pomocí dělení a násobení bodového odhadu $\exp(b)$.

Přímo v této práci jsou prezentovány buď AFT modely tam, kde test neověřil statisticky významný rozdíl mezi modelem s kategoriální a kvantitativní proměnnou, nebo grafické znázornění odhadů parametrů β s 95% intervalem spolehlivosti. V tomto znázornění vyznačuje modrá barva období před krizí, červená barva období během krize a zelená barva období po krizi. Shora dolu jsou vždy znázorněny odhady za soubory 1, 2 a 3 dohromady. Veškeré v textu zmíněné odhady procentuálního rozdílu pochází z přílohy B.

4.3.1 Regiony

Doba do nalezení práce v souboru 1 se lišila podle regionů pouze v předkrizovém a krizovém období, v pokrizovém období se rozdíly zmenšily natolik, že nebyl nalezen statisticky významný rozdíl v této proměnné. Poslední sloupec tabulky A1 v příloze ukazuje, že statisticky významné změny proběhly v regionech Severozápad, kde doba do nalezení práce průběžně zkracovala, Jihozápad, kde se postupně prodlužovala, a Severovýchod, kde se prodloužila v období krize a zůstala přibližně stejná v období po krizi.

Obrázek 4.19 znázorňuje odhady modelu pro předkrizové a krizové období. V předkrizovém období ukazuje rozdělení podle doby do nalezení práce na tři skupiny regionů - Praha, Jihozápad, Severovýchod, Střední Morava a Jihovýchod s nejkratší dobou, Střední

Čechy a Moravskoslezsko s dobou do nalezení práce o cca 31 % vyšší, a Severozápad s dobou do nalezení práce vyšší o cca 164 %. AFT model pro krizové období ukazuje zmíněné srovnání těchto dob, kdy vybočuje jen region Severozápad s dobou do nalezení práce o cca 50 % delší než ve zbytku republiky.

V souboru 2 pozorujeme podobný vývoj ve vztahu doby nezaměstnanosti k regionu jako u předchozích dat, tedy vzájemné přiblížení. V předkrizovém období byly vidět celkem 4 skupiny s navzájem odlišnou dobou nezaměstnanosti, v krizovém 3 a v pokrizovém 2, viz obrázek 4.19.

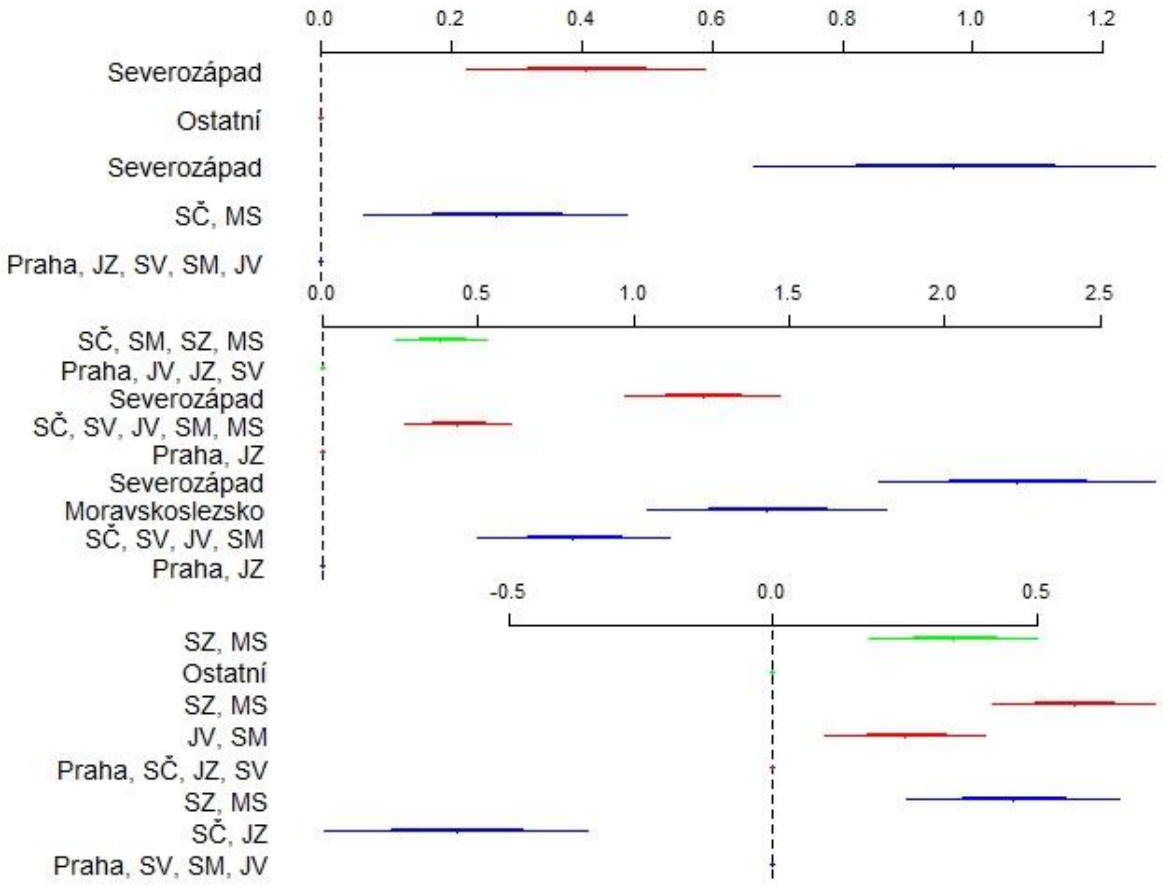
První skupinu tvořila Praha a Jihozápad, ve druhé skupině se sdružily regiony Střední Čechy, Severovýchod, Jihovýchod a Střední Morava – doba nezaměstnanosti byla oproti první skupině vyšší o přibližně 124 %. Osamoceně pak stojí regiony Moravskoslezsko s dobou nezaměstnanosti vyšší o zhruba 317 % a Severozápad s nejdelší dobou nezaměstnanosti, oproti první skupině vyšší o 831 %.

Během krize se tyto rozdíly snížily, přičemž Moravskoslezsko se nelišilo statisticky významně od druhé skupiny. Ta měla oproti první skupině dobu nalezení práce delší o zhruba 46 % a Severozápad o asi 86 %.

Po krizi došlo k dalšímu sblížení dob nezaměstnanosti v různých krajích, které utvořily dvě vzájemně odlišné skupiny. K Praze s Jihozápadem se přidružily regiony Severovýchod a Jihovýchod, Severozápad se naopak dostal na úroveň s regiony Střední Čechy, Střední Morava a Moravskoslezsko. Doba do nalezení práce zde byla oproti první skupině delší o zhruba 46 %.

Z neparametrických testů (tabulka A4 v příloze) pak vyplývá, že nejvyšší pohyby nastaly v regionech Severozápad a Moravskoslezsko, ve kterých se doba nezaměstnanosti průběžně zkracovala. V ostatních regionech nebyly pohyby funkce přežití statisticky významné.

Podobně jako v předchozích dvou případech je i u dat ze souboru 3 vidět jisté srovnání doby nezaměstnanosti mezi jednotlivými regiony. Před krizí se daly rozlišit tři skupiny – první s regiony Praha, Severovýchod, Střední Morava a Jihovýchod, druhá s regiony Střední Čechy a Jihozápad a třetí s regiony Moravskoslezsko a Severozápad. V předkrizovém období byla doba nezaměstnanosti ve druhé skupině oproti první kratší o 45 % a ve třetí vyšší o 57 %. Kumulovaně pak třetí skupina měla vyšší dobu nezaměstnanosti oproti druhé o zhruba 185 %.



Obrázek 4.19 Odhady parametrů podle regionu

V krizovém období se skupiny odlišily trochu jinak – v první byly vedle Prahy Střední Čechy, Jihozápad a Severovýchod, ve druhé Střední Morava a Jihovýchod, třetí zůstala stejná. Druhá skupina měla oproti první dobu nezaměstnanosti vyšší o 22 % a třetí o 70 %

V pokrizovém období byla doba nezaměstnanosti odlišná v regionech Severozápad a Moravskoslezsko – oproti zbytku republiky vyšší o zhruba 40 %.

Statisticky významné pohyby (tabulka A8 přílohy) byly u regionů Praha, ve kterém se doba nezaměstnanosti zkrátila v období krize a pak se udržela, v regionu Střední Čechy, kde vzrostla v pokrizovém období, v regionu Jihozápad, kde se průběžně prodlužovala, a v regionu Moravskoslezskou, kde se průběžně zkracovala.

Všechny výsledky poukazují na sbližování se doby do nalezení zaměstnání, respektive doby nezaměstnanosti mezi jednotlivými regiony v průběhu času. V souborech 2 a 3 se dá usuzovat na efekt nových nezaměstnaných, kdy vyšší míra propouštění ve více postižených regionech může snížit průběh doby nezaměstnanosti, nicméně v souboru 1 by se tento efekt projevovat neměl, takže lze usuzovat spíše na obecné zlepšení podmínek na trhu práce v těch

regionech s obecně nejdelší dobou nezaměstnanosti, zejména v regionech Severozápad a Moravskoslezsko.

4.3.2 Počet osob v domácnosti

V souboru 1 existuje statisticky významný rozdíl podle počtu osob žijících v domácnosti daného člověka pouze v předkrizovém období. Podle testu pomocí věrohodnostního poměru nebyl zjištěn rozdíl mezi modelem s kvantitativní proměnnou a její kategorizací.

Z modelu v tabulce 4.24 vyplývá, že za každého člena domácnosti klesá doba do nalezení práce o necelých 9 %, přičemž v domácnosti s 0 členy je v modelu mediánová doba do nalezení práce 8,81 měsíců.

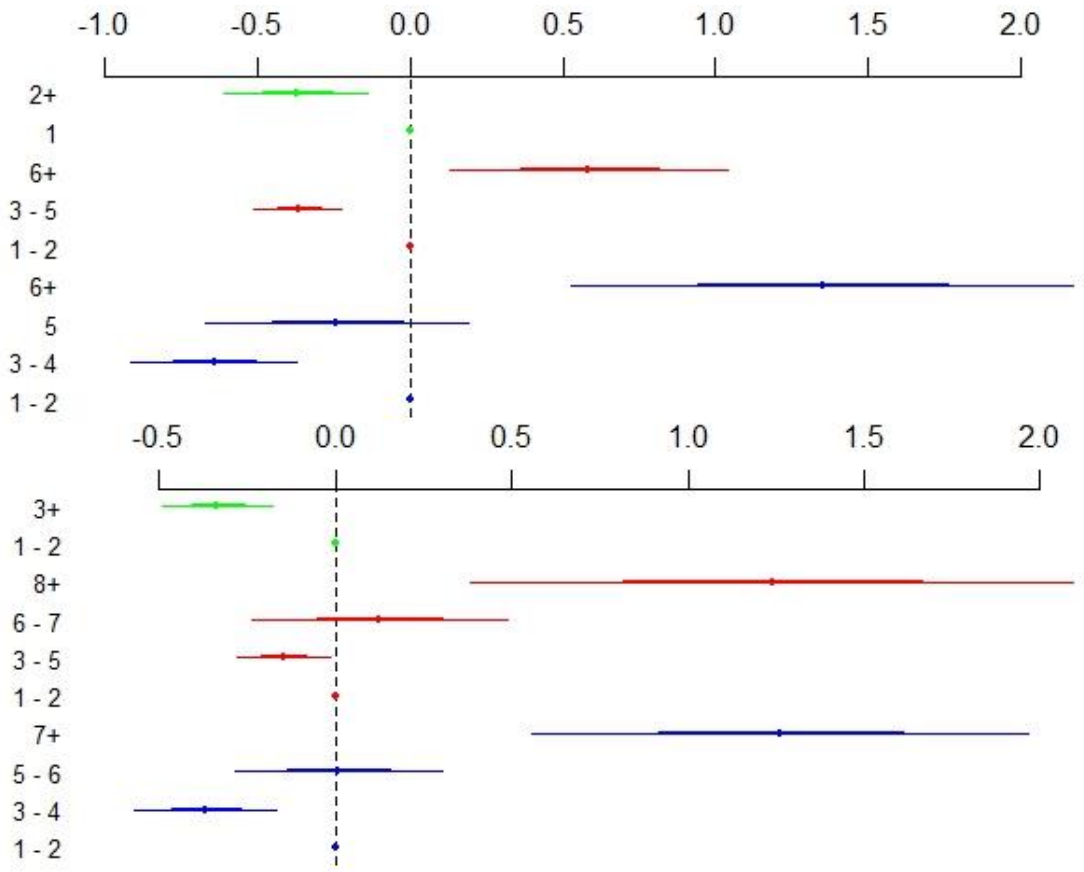
Nebyla zaznamenána statisticky významná změna pro jednotlivé hodnoty. (tabulka A11 přílohy)

Tabulka 4.24 AFT model podle počtu členů domácnosti, soubor 1

pocod během krize	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		8,81	0,66	25,35	2,05	13,34	5,32
Pocod	0,912	1,045						

V souboru 2 vypadá vztah počtu členů domácnosti k době nezaměstnanosti trochu složitěji, což zachycuje obrázek 4.20. S rostoucím počtem členů domácnosti nejdříve doba nezaměstnanosti klesá a následně roste. V předkrizovém období je nejkratší u domácností se 3 až 4 členy, pak s 5 členy, 1 členem a nejvyšší v domácnostech s více než s 6 členy. V krizovém období je nejnižší nezaměstnanost u domácností se 3 až 5 členy, vyšší pak u jednočlenných domácností a nejvyšší taktéž u více než šestičlenných. Pokrizové období se vymyká, v něm se doma nezaměstnanosti statisticky významně neliší od 2 členů výš a nejvyšší je tak u jednočlenných domácností.

Měnila se doba nezaměstnanosti u dvoučlenných domácností, ve kterých klesla v krizovém období oproti předkrizovému, a u pětičlenných domácností, kde došlo k podobnému vývoji (tabulka A14 přílohy). To se odráží ve slučování jednotlivých hodnot.



Obrázek 4.20 Odhady parametrů podle počtu členů domácností, soubory 2 a 3.

V souboru 3 vychází obdobné sloučení jako v souboru 2 – před krizí byla nejkratší doba nezaměstnanosti u lidí z domácností se 3 či 4 členy, u lidí z domácností s 5 až 6 lidmi nebyla statisticky významně odlišná od lidí v domácnostech s 1 nebo 2 členy (tabulka A19 přílohy). V krizi byla nejnižší doba nezaměstnanosti u lidí z domácností se 3 až 5 členy, u lidí z domácností se 6 nebo 7 členy nebyla statisticky významně odlišná od lidí z domácností s 1 nebo 2 členy (tabulka A20 přílohy). Po krizi došlo ke kategorii počtu osob v domácnostech s více než 2 členy a nejnižší doba nezaměstnanosti byla u lidí z domácností s 1 nebo 2 členy.

Doba nezaměstnanosti se statisticky významně měnila u nezaměstnaných lidí z domácností s 1 a 2 členy, u kterých nejdříve došlo k poklesu v období krize a následně mírnému nárůstu v období po krizi. U domácností s více než 6 členy pak doba nezaměstnanosti průběžně klesala (tabulka A18 přílohy).

Obecně nejkratší dobu do nalezení zaměstnání či dobu nezaměstnanosti mají lidé z domácnosti se 3 až 4 členy. Dá se zde usuzovat na vliv dalších vlastností těchto osob, protože se může typicky jednat o lidi z rodin v produktivním věku. Vysoký počet osob v domácnosti zase může značit slabší sociální status a tedy souhrn faktorů zhoršujících výhledy na trhu práce, zejména nižší vzdělání.

4.3.3 Vztah k osobě v čele domácnosti

V souboru 1 měli statisticky významně kratší dobu do nalezení práce děti osob v čele domácnosti oproti jiným členům domácností, a to před krizí a během krize o 27 respektive 18 %, přičemž v období krize byli do jedné skupiny s dětmi zařazeni i ostatní členové domácnosti, což je znázorněno v obrázku 4.21. V období po krizi již statisticky významný rozdíl nalezen nebyl.

Testy v tabulce A22 přílohy taky ukazují skutečnost, že u dětí se statisticky významně průběžně prodloužila doba do nalezení práce, což je v souladu s těmito výsledky.

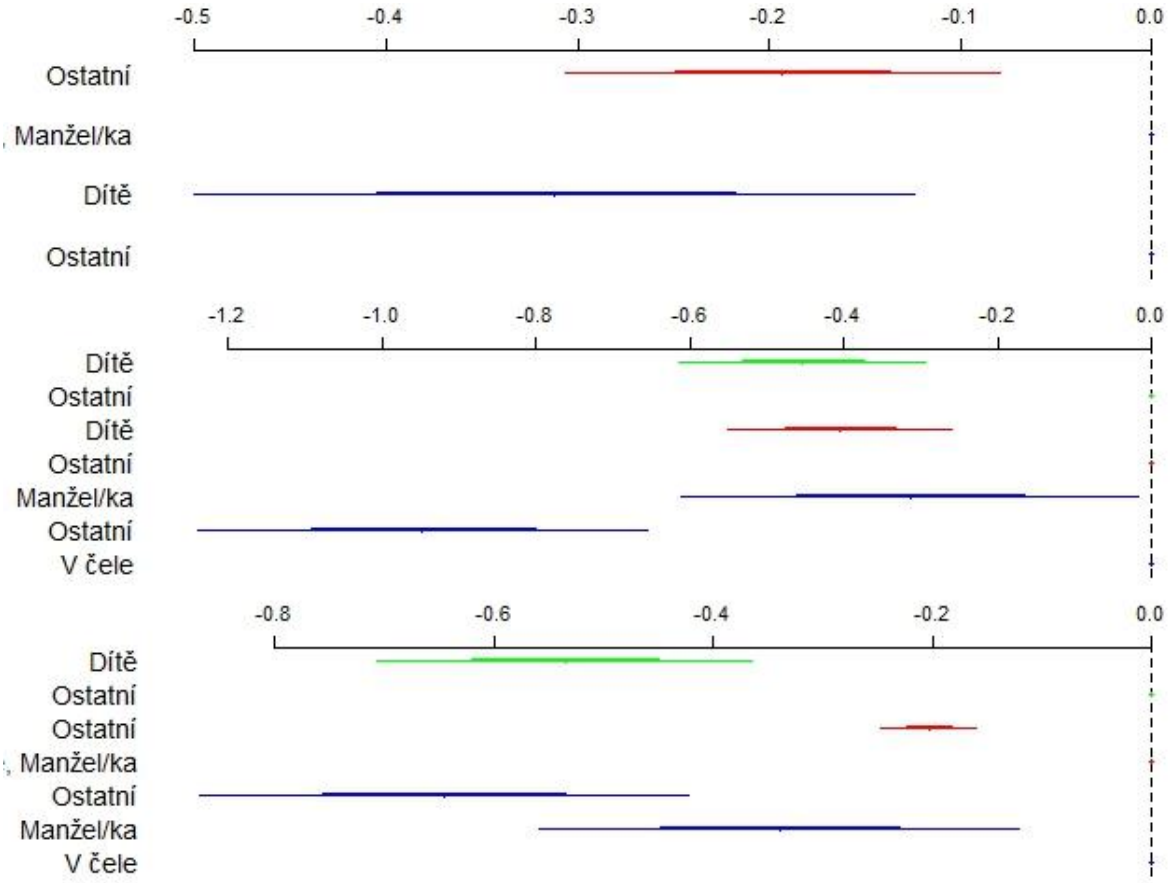
V souboru 2 vypadají výsledky odlišně, nejkratší dobu do nalezení práce mají děti osob v čele domácnosti ve všech třech obdobích (postupně o 62 %, 33 % a 36 % oproti referenční kategorii). V období před krizí byli s dětmi sloučeni ještě další členové domácnosti a statisticky významně se mezi touto skupinou a osobami v čele domácnosti umístili manželé/manželky osob v čele domácnosti (o 27 % kratší doba do nalezení práce).

Tabulka A25 přílohy ukazuje, že statisticky významný pohyb doby do nalezení práce nastal u osob v čele domácnosti, u kterých po krizi výrazně poklesla, a u manželů/manželek, u kterých se zkracovala postupně přes obě období.

V souboru 3 je doba nezaměstnanosti nejkratší opět u dětí, postupně o 48 %, 18 % a 42 % oproti referenční kategorii. Před krizí jsou s dětmi spojeni ostatní členové domácnosti a samostatně vystupují manželé/manželky obdobně jako u souboru 2, jejich doba nezaměstnanosti byla o 29 % kratší než u osob v čele domácnosti. Během krize byla doba nezaměstnanosti ostatních členů domácnosti blíže k dětem než k osobám v čele domácnosti a jejich manželům/manželkám, a po krizi jsou již děti jako kategorie osamostatněné.

Podle tabulky A29 přílohy byly pohyby doby nezaměstnanosti stejného směru jako u souboru 2.

Nejvýraznější rozdíly jsou mezi dětmi osob v čele domácností, kteří mají stabilně nejkratší dobu do nalezení práce či dobu nezaměstnanosti. Toto může být způsobeno zejména jejich věkem, jak bude uvedeno dále, nejmladší osoby mívají nejkratší dobu nezaměstnanosti. V některých případech je zajímavé postavení manžela či manželky osoby v čele domácnosti, které nachází práci o něco rychleji. Pro tento jev mě žádné vysvětlení nenapadá.

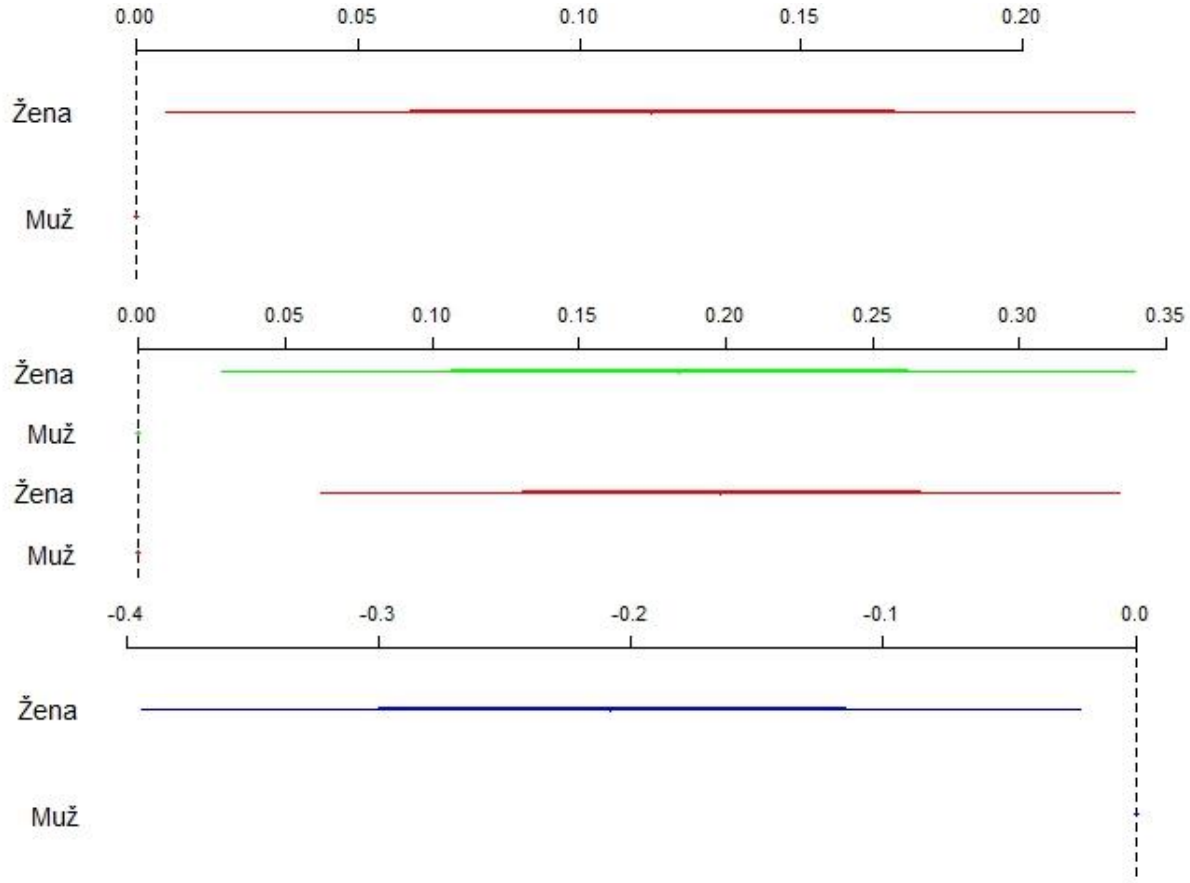


Obrázek 4.21 Odhady parametrů podle postavení v domácnosti

4.3.4 Pohlaví

V souboru 1 byly rozdíly v době do nalezení práce mezi pohlavími relativně nízké. Neparametrické testy (viz tabulka A33 přílohy) vycházely statisticky významně, ale u parametrického AFT modelu vyšel rozdíl významný jen v období během krize. Ženy v něm měly dle obrázku 4.22 dobu do nalezení práce delší o zhruba 11 %. Posuny mezi jednotlivými obdobími nebyly ani pro jedno pohlaví statisticky významné (tabulka A33).

V souboru 2 byly statisticky významné rozdíly podle pohlaví u období během krize a po krizi, o 22 % respektive o 20 %.



Obrázek 4.22 Odhady parametrů podle pohlaví

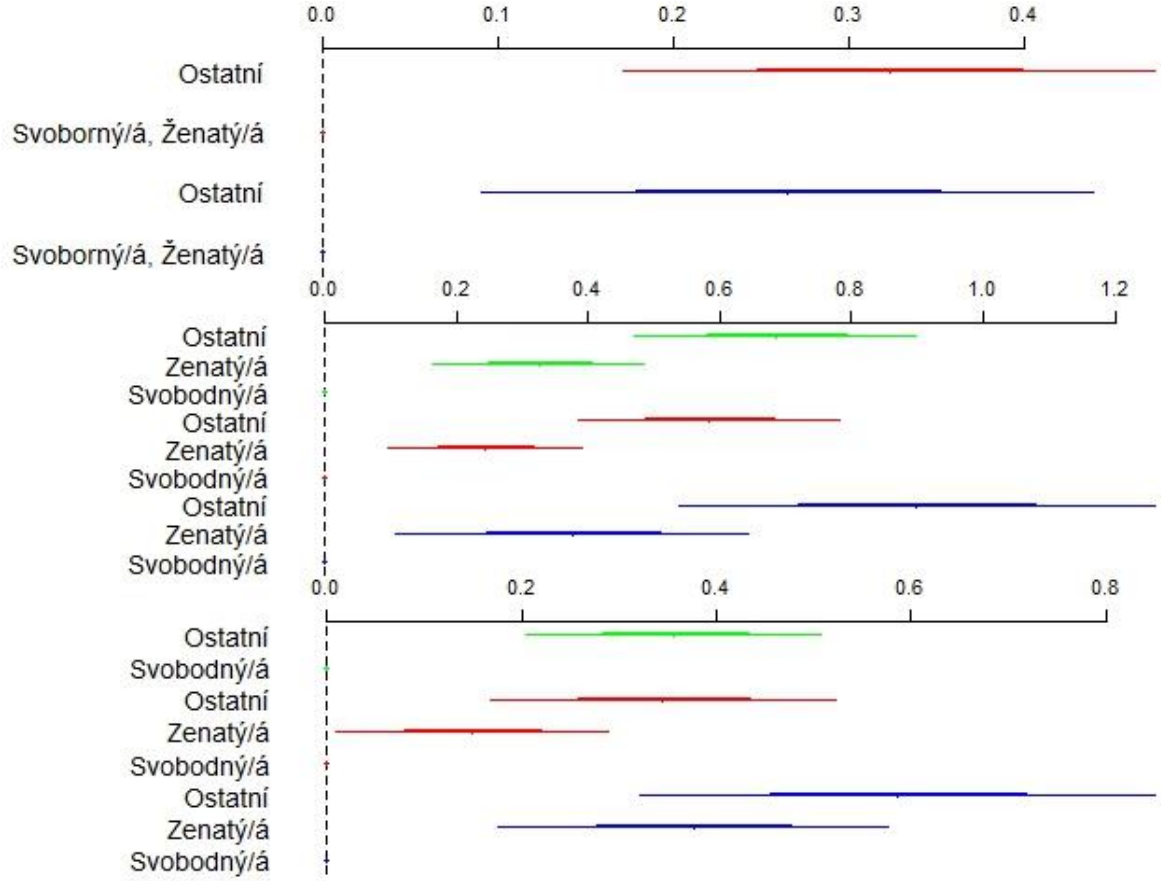
V souboru 3 vyšel statisticky významný rozdíl v době nezaměstnanosti naopak v období před krizí, kdy ženy nacházely práci o zhruba 19 % dříve. Tento výsledek zaznamenaný taktéž v obrázku 4.22 je v rozporu s ostatními, přičemž v neparametrickém testu vyšel rozdíl statisticky nevýznamný (viz tabulka A38 přílohy).

Celkově se dá říci, že rozdíly mezi pohlavími nejsou příliš velké, přičemž při 1% hladině významnosti by se za významný prohlásil pouze rozdíl v období krize v souboru 2.

4.3.5 Rodinný stav

Doba do nalezení práce podle rodinného stavu se u prvního souboru statisticky významně lišila před krizí a během krize. Rozvedení s ovdovělými měli oproti svobodným se ženatými dobu do nalezení práce delší o 30 % respektive o 38 %, což je zaznamenáno v obrázku 4.23. V období po krizi byl rozdíl statisticky nevýznamný v AFT modelu, v log-rank testu byl významný těsně (viz tabulka A40 přílohy).

Statisticky významný rozdíl mezi jednotlivými obdobími byl zaznamenán jen u svobodných, a to též velmi těsně, viz tabulka A40 přílohy. Zdá se, že u nich doba do nalezení práce postupně stoupala.



Obrázek 4.23 Odhady parametrů podle rodinného stavu

V souboru 2 byly zaznamenány výraznější rozdíly, které shrnuje obrázek 4.23. Před krizí, během krize i po krizi bylo stejné rozdělení i pořadí z hlediska doby nezaměstnanosti, kdy nejkratší byla u svobodných, poté u ženatých/vdaných a nejdelší u rozvedených a ovdovělých, kteří vytvořili samostatnou skupinu. Nejvyšší rozdíly byly v období před krizí – o 46 % a o 146 % delší doba do nalezení práce u ženatých/vdaných respektive u rozvedených a ovdovělých. Během krize byly tyto rozdíly naopak nejkratší – 28 a 79 %. Po krizi zase rozdíly mírně narostly na 38 a 98 %.

Statisticky významné pohyby byly zaznamenány u ženatých/vdaných a rozvedených, u kterých doba do nalezení práce klesla v období krize a zůstala na podobné úrovni i po krizi (tabulka A43).

Pro soubor 3 vznikly podobné modely v období před krizí a během krize jako u druhého souboru – relativní rozdíly byly akorát o něco nižší, tedy přesněji o 46 % a 80 % před krizí a o 16 % a 41 % během krize. Po krizi již nebyl statisticky významný rozdíl mezi ženatými/vdanými a rozvedenými a ovdovělými. Tyto tři skupiny dohromady měly dobu nezaměstnanosti zhruba o 43 % vyšší než svobodní. Výsledky jsou v tabulce 4.34.

Statisticky významný pohyb doby nezaměstnanosti byl zaznamenán u ženatých/vdaných, kdy v souladu s výše napsaným doba nezaměstnanosti stoupla v období po krizi (viz tabulka A47 přílohy).

Obecně nejkratší dobu do nalezení zaměstnání i dobu nezaměstnanosti vykazují svobodní lidé a nejdelší ovdovělí a rozvedení, což může být opět zapříčiněno vlivem věku a ne samotného rodinného stavu.

4.3.6 Věk

V souboru 1 vycházely statisticky významně jak model počítající s věkem jako kvantitativní vysvětlující proměnnou tak i model počítající s věkovými skupinami jako kategoriální proměnnou. Model s kategoriální proměnnou měl nicméně statisticky významně vyšší věrohodnost, proto dostal přednost.

V souboru 1 byla nejkratší doba do nalezení práce u nejmladších a nejstarších lidí (15 – 29, 15 – 34 respektive 15 – 19 na jedné straně a 60+ respektive 65+ na druhé straně) a lidé středního staršího věku (30 – 59, 34 – 59 respektive 30 – 64) měli dobu do nalezení práce nejdelší. Ačkoliv bodové odhady v obrázku 4.24 naznačují kratší dobu do nalezení práce u nejstarších, v AFT modelu nejsou příslušné parametry statisticky významné, tato doba tedy není statisticky významně odlišná od doby do nalezení práce u nejmladších (tabulky A52 až A54 přílohy).

Statisticky významně se lišily hodnoty přes různá období jen u nejmladší kategorie věku 15 – 19 let, u které doba do nalezení práce vzrostla v období krize a následně zase klesla na předkrizovou hodnotu, což je zaznamenáno v tabulce A51 přílohy.

V souboru 2 se modely s kvantitativním a kategoriálním pojetím věku statisticky významně nelišily, proto dostal přednost model s věkem jako kvantitativní proměnnou. To znamená, že doba do nalezení práce roste s věkem, pro tři sledovaná období za každý rok věku o 4,6 %, 2,5 % a 3,5 %, jak ukazuje tabulka 4.25.

Z výsledků v tabulce A55 přílohy vyplývá, že doba do nalezení práce se statisticky významně měnila u mladších kategorií, nejvýrazněji u kategorie 15 – 19 let, kde v průběhu krize doba do nalezení práce poklesla a následně zase narostla. Obdobný průběh lze vysledovat u kategorií 20 – 24 let a 25 – 29 let.

Tabulka 4.25 AFT model podle věku, soubor 2

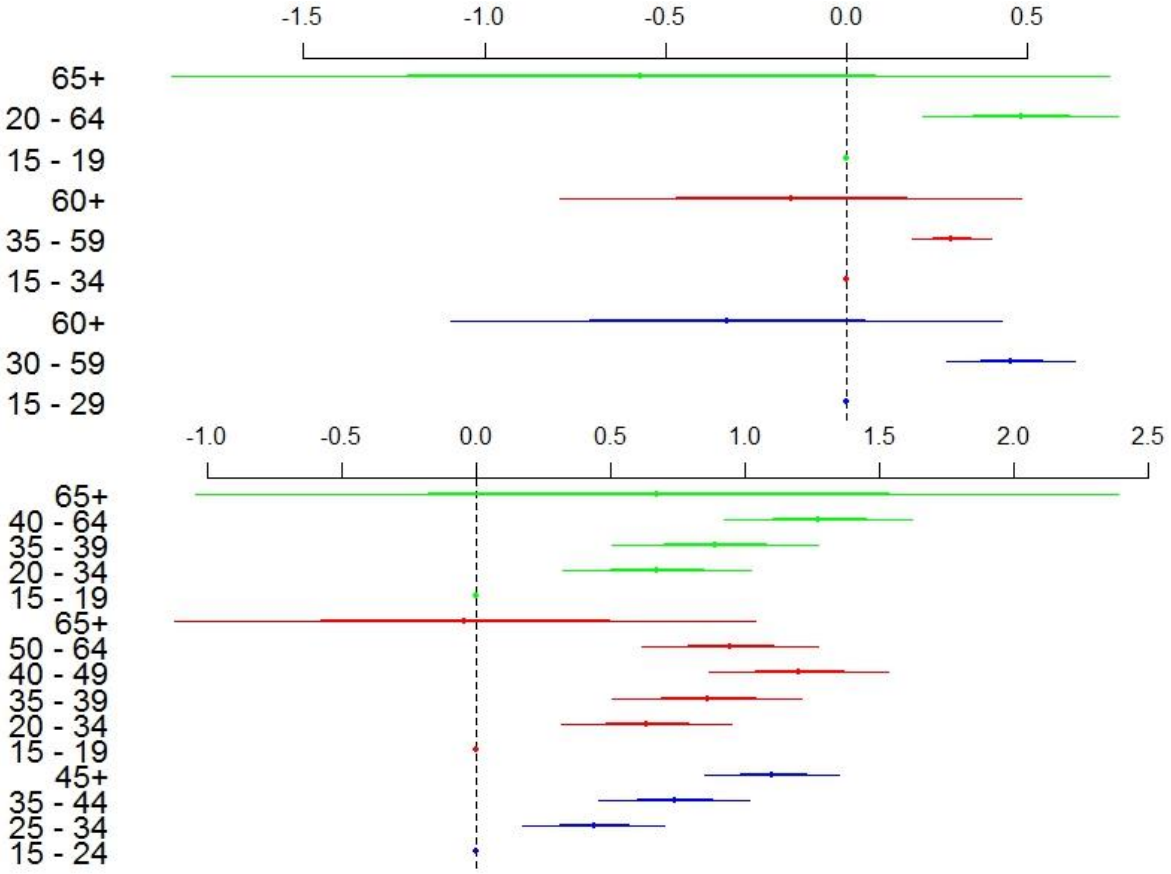
Vek před krizí	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		11,19	2,12	152,06	34,19	-75,07	9,65
Vek	1,046	1,010						
Vek během krize	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		12,15	1,25	82,19	8,97	83,75	9,17
Vek	1,025	1,005						
Vek po krizi	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		9,52	1,05	69,07	8,80	89,11	7,34
Vek	1,035	1,006						

V rámci souboru 3 lze pro období před krizí vysledovat, že s rostoucím věkem se prodlužuje doba nezaměstnanosti. Nejnižší je u kategorie 15 – 24 let, o zhruba 55 % vyšší u kategorie 25 – 34 let, o 109 % u kategorie 35 – 44 let a o 200 % vyšší u lidí ve věku 45 let a vyšším.

V krizovém období lze u doby nezaměstnanosti vidět naopak zkracující se dobu nezaměstnanosti u starších věkových skupin. Ve srovnání s nejkratší dobou nezaměstnanosti u skupiny 15 – 19 let je ve skupině 20 – 34 let doba nezaměstnanosti vyšší o 88 %, ve skupině 35 – 39 let o 136 %, ve skupině 40 – 49 let o 230 % a pak klesá; ve skupině 50 – 64 let je doba nezaměstnanosti ve srovnání s referenční kategorií vyšší o 157 % a ve skupině nad 64 let je vyšší o 96 %.

Po krizi vypadá rozdělení skupin podobně jako v krizovém období, ale kategorie 40 – 49 let a 50 – 64 let již nebyly statisticky významně odlišné a tak byly sloučeny. Ve srovnání s nejmladšími jsou doby nezaměstnanosti delší postupně o 95 %, 143 %, 256 % a 96 %. Lze tedy hovořit o tom, že rozdíly v době nezaměstnanosti byly podobné napříč obdobími.

Statisticky významné změny v době nezaměstnanosti se udály v tomto souboru u starších kategorií, specificky 50 – 54 let a 55 – 59 let, ve kterých v období krize doba nezaměstnanosti klesla a po krizi se zase vrátila na předkrizovou hodnotu, viz tabulka A59 přílohy.



Obrázek 4.24 Odhady parametrů podle věku

Obecně se doba do nalezení práce či doba nezaměstnanosti s věkem prodlužuje s výjimkou nejstarších věkových kategorií. Tato výjimka může být způsobena návratem z ekonomické neaktivity, kdy dotyčná osoba začne aktivně hledat práci a stává se tak znovu ekonomicky aktivním, přičemž předchozí doba ekonomické neaktivity se do doby nezaměstnanosti nezapočítává.

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.

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. Prodloužení doby nezaměstnanosti v krizi u nejmladších osob v souboru 1 (tedy souboru těch, co práci skutečně nalezli) je pak nejspíš důsledkem zhoršení podmínek na trhu práce, na kterém se nejmladší neuplatňovali tak rychle jako před krizí a po krizi.

4.3.7 Země narození

V souboru 1 ani souboru 2 nebyl zaznamenán statisticky významný rozdíl v době do jejího nalezení u lidí podle země narození. Roli zde nejspíš hraje nedostatek pozorování, zejména lidí s jiným místem narození než Česká republika

V souboru 3 byl nalezen statisticky významný rozdíl pro období před krizí, kdy lidé narození mimo Českou republiky měli dobu nezaměstnanosti delší o zhruba 93 %, odhad je v tabulce 4.26. Pro další období nebyl zaznamenán statisticky významný rozdíl. Výsledky můžou být důsledkem relativně malého zastoupení lidí s odlišnou zemí narození.

Tabulka 4.26 AFT model podle země narození, soubor 3

Zemnar před krizí	$\exp(x^T b)$	$\exp(\text{s.e.}*u_{0,975})$	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
ČR	1,000		10,15	0,48	74,82	5,44	103,19	7,85
Jiná	1,927	1,609	19,56	4,66	144,23	35,30	198,91	15,14

4.3.8 Státní příslušnost

V souboru 1 byla státní příslušnost statisticky významným faktorem pouze v období po krizi, ve kterém lidé s jinou, než českou státní příslušností měli dobu do nalezení zaměstnání delší o zhruba 81 %, jak je uvedeno v tabulce 4.27. Tento vztah ale nebyl statisticky významný při použití neparametrického testu, viz tabulka A67 přílohy. Opět zde můžeme očekávat vliv nedostatečného počtu pozorování.

Tabulka 4.27 AFT model podle státní příslušnosti, soubor 1

Stprisl po krizi	$\exp(x^T b)$	$\exp(\text{s.e.}*u_{0,975})$	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
ČR	1,000		6,52	0,19	19,76	0,86	10,34	2,55
Jiná	1,811	1,467	11,81	2,28	35,79	7,00	18,72	4,63

V souboru 2 ani 3 nebyl nalezen statisticky významný vztah doby do nalezení práce a státní příslušnosti. Opět může být na vině datová nedostatečnost.

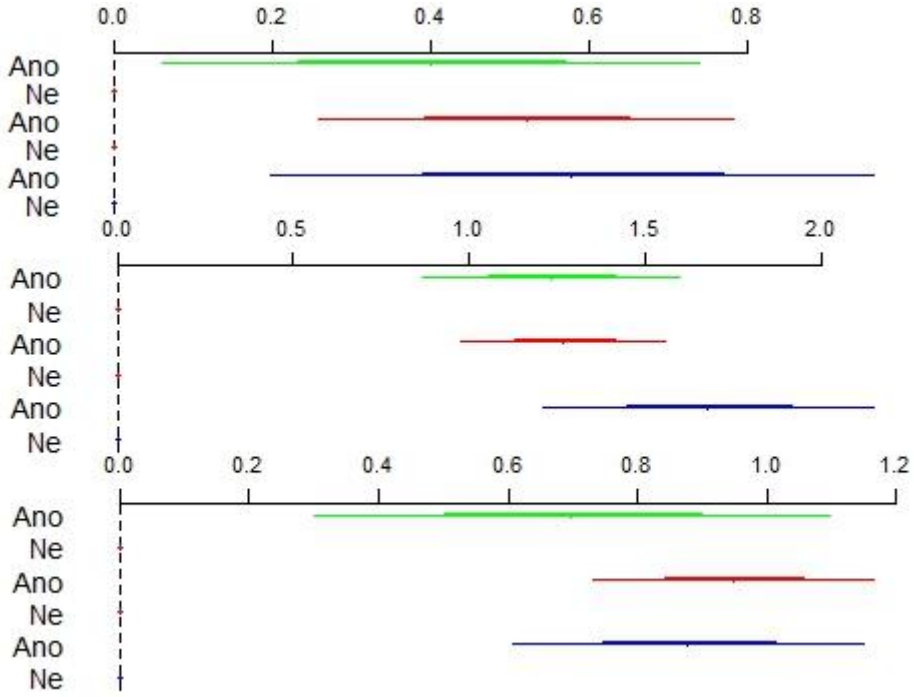
4.3.9 Zdravotní postižení

Lidé se zdravotním postižením mají dobu do nalezení zaměstnání statisticky významně delší než lidé bez něj, což je konzistentní zjištění napříč všemi obdobími i soubory.

V souboru 1 byla tato doba delší postupně o 78 %, 68 % a 49 %, což je vyjádřeno v obrázku 4.25. U lidí se zdravotním postižením nedocházelo ke statisticky významnému posunu doby nezaměstnanosti napříč obdobími.

V souboru 2 byla doba do nalezení zaměstnání delší u postižených postupně o 436 %, 356 % a 343 % oproti lidem bez zdravotního postižení. U zdravotně postižených ani zde nedošlo ke statisticky významným posunům doby nezaměstnanosti.

V souboru 3 byla doba nezaměstnanosti u postižených delší postupně o 141 %, 156 % a 101 %. Stejně jako v předchozích případech se doba nezaměstnanosti u postižených v průběhu doby statisticky významně nezměnila.



Obrázek 4.25 Odhady parametrů podle zdravotního postižení

Doba do nalezení práce u postižených konzistentně vykazuje, že během krize a po krizi došlo k vyrovnání doby nezaměstnanosti mezi nimi a zdravými, k čemuž zřejmě přispěl především posun doby nezaměstnanosti u zdravých lidí, a ne u postižených.

4.3.10 Vzdělání

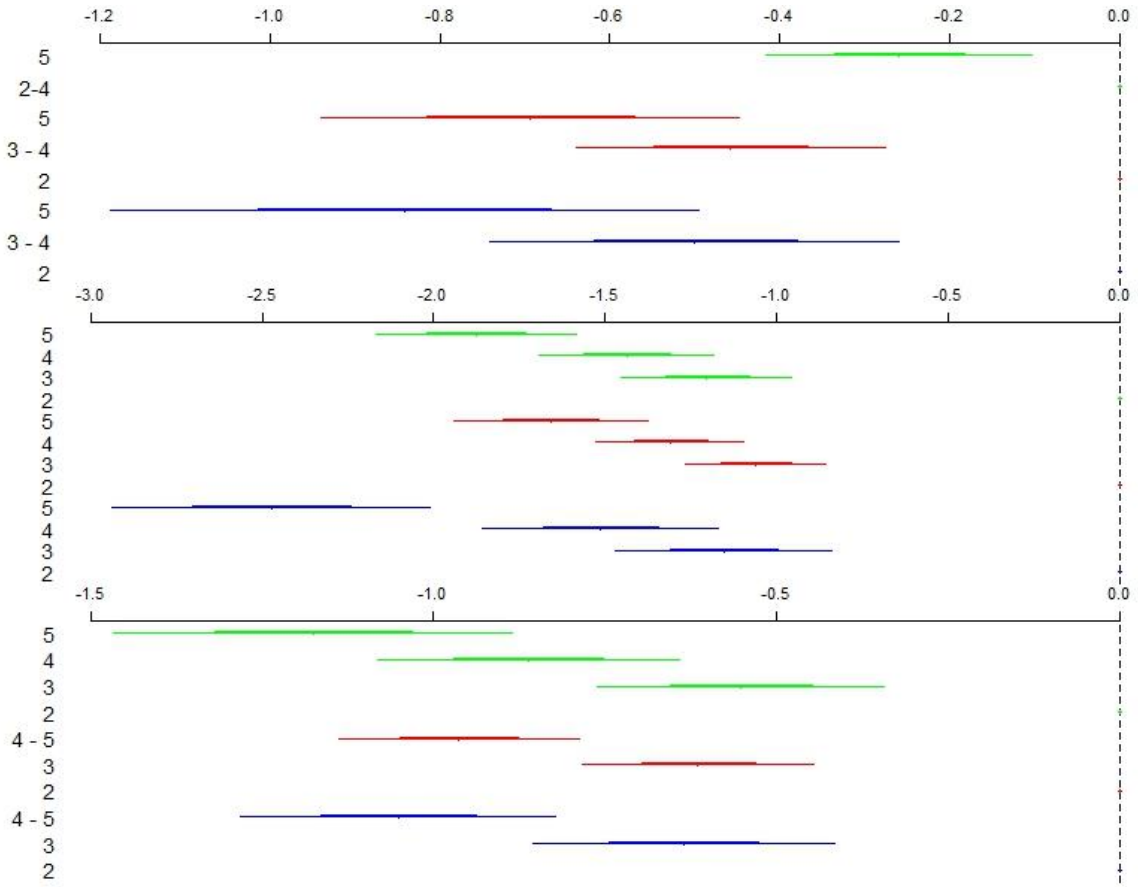
Doba do nalezení práce v souboru 1 se postupně sbližovala. V období před krizí a během krize hledali práci nejdéle lidé bez ukončeného středoškolského vzdělání (ISCED = 2). Oproti nim nacházeli práci lidé s ukončeným středoškolským vzděláním práci dříve o cca 39 % respektive o 37 %, a lidé s vysokoškolským vzděláním o 57 % a 50 % rychleji. V období po krizi nebyl zjištěn statisticky významný rozdíl v době do nalezení práce mezi lidmi se základním a ukončeným středoškolským vzděláním. Lidé s univerzitním vzděláním nacházeli práci o zhruba 23 % dříve. Výsledky jsou v obrázku 4.26.

Podle výsledků v tabulce A82 přílohy se doba nezaměstnanosti staticky významně posunula právě u lidí se středoškolským vzděláním, a to směrem nahoru v období po krizi. Toto zjištění je konzistentní s výsledky AFT modelů.

Doba do nalezení zaměstnání v souboru 2 se statisticky lišila mezi všemi kategoriemi vzdělání, přičemž oproti základnímu vzdělání byla u středního bez maturity kratší postupně o 68 %, 65 % a 70 %, u středního s maturitou o 78 %, 73 % a 76 % a u vysokoškolského vzdělání o 92 %, 81 % a 85 %.

Tabulka A86 přílohy ukazuje, že statisticky významný pokles doby nezaměstnanosti nastal u lidí se středoškolským vzděláním bez maturity, nicméně při přísnější 1% hladině významnosti by už tato změna nebyla statisticky významná. Rozdíly se zdají být stabilní napříč obdobími.

V souboru 3 nebyl v období před krizí a během krize statisticky významný rozdíl v době nezaměstnanosti mezi lidmi s maturitou a vysokou školou. Nejdelší dobu nezaměstnanosti měli lidé se základním vzděláním, oproti nim lidé se střední školou bez maturity měli dobu nezaměstnanosti kratší o 47 % a 46 %, lidé s maturitou nebo vysokou školou pak o 65 % respektive 62 %. V období po krizi už se statisticky významně odlišuje doba nezaměstnanosti u lidí s maturitou oproti lidem s vysokou školou. Oproti lidem se základním vzděláním je doba nezaměstnanosti kratší postupně o 42 %, 58 % a 69 %.

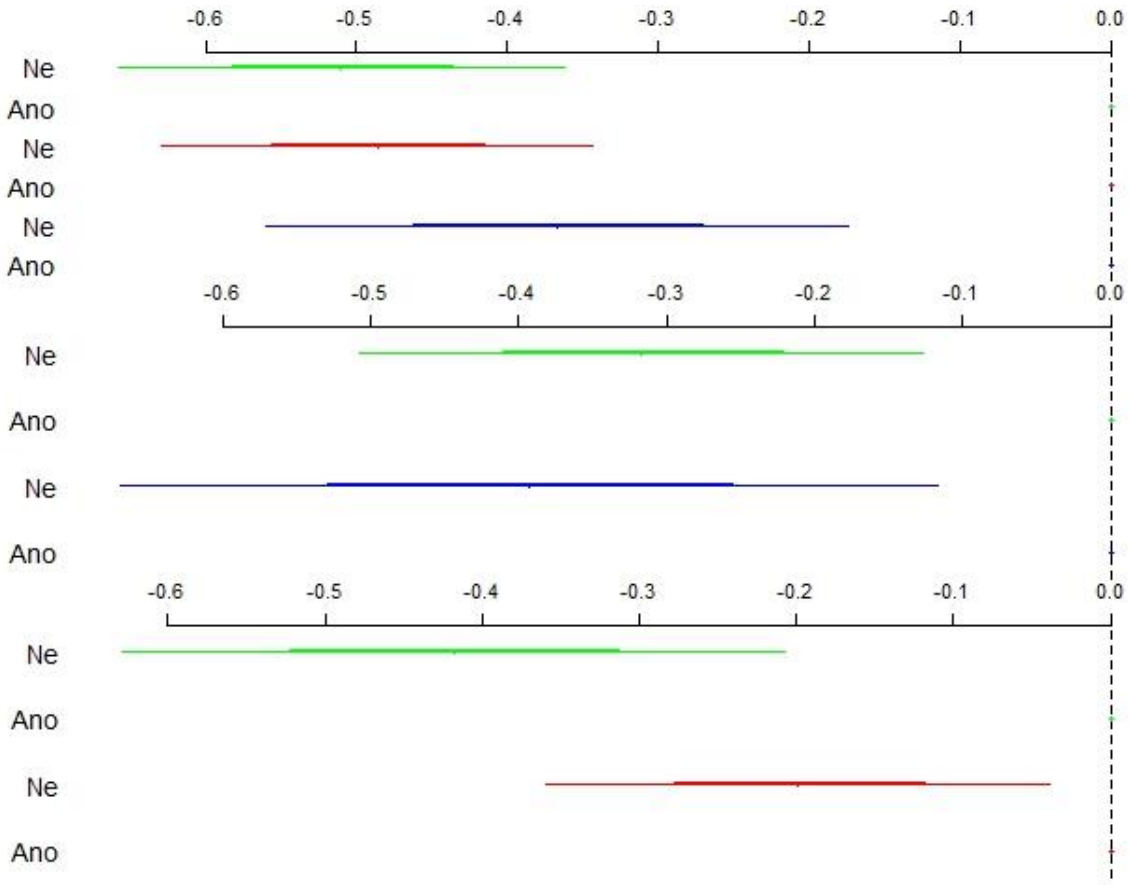


Obrázek 4.26 Odhady parametrů podle stupně vzdělání

4.3.11 Registrace na Úřadu práce

V souboru 1 měli lidé neregistrovaní na Úřadu práce kratší dobu do nalezení práce, a to ve všech sledovaných obdobích. Doba do nalezení zaměstnání byla u neregistrovaných kratší před krizí o 31 %, během krize o 38 % a po krizi o 40 %, přičemž nebyl zaznamenán statisticky významný rozdíl v rámci různých období ani u lidí registrovaných ani u lidí neregistrovaných (tabulka A94 přílohy).

V souboru 2 byla doba do nalezení práce statisticky významně odlišná v období před krizí a po krizi, ale ne během krize. U neregistrovaných byla kratší o 32 % respektive o 27 %. Doba do nalezení práce se statisticky významně změnila u registrovaných, u kterých postupně klesala, ale ne u neregistrovaných (tabulka A98 přílohy).



Obrázek 4.27 Odhady parametrů podle registrace na ÚP

V souboru 3 byla doba nezaměstnanosti během krize kratší u neregistrovaných o 18 % a v období po krizi o 34 %. Před krizí nebyl rozdíl statisticky významný. Podle log-rank testu nebyl statisticky významný ani rozdíl v období během krize (tabulka A101 přílohy).

K posunu došlo podobně jako u všech nezaměstnaných, kdy doba nezaměstnanosti registrovaných na Úřadu práce nejdříve klesla během krize a následně stoupla. U neregistrovaných byl statisticky nevýznamný, ale poměrně těsně. U nich nejdříve doba nezaměstnanosti klesla poté si udržela nižší úroveň, což je konzistentní s výsledky výše, viz tabulka A101 přílohy.

Zvláštností těchto výsledků je fakt, že registrovaní na ÚP mají obecně delší dobu nezaměstnanosti. Tento fakt může být vysvětlen obrácenou kauzalitou, tedy můžeme uvažovat o tom, že lidé, kteří nenachází práci dost rychle, přichází na úřad práce. Tuto tezi by podporovaly skutečnosti uvedené v kapitole 1.3 a tabulce 1.1, ve kterých dle srovnání veřejně dostupných údajů jsou lidé registrovaní na ÚP nezaměstnaní *kratší* dobu. Tento paradox může být vysvětlen právě pozdní registrací na ÚP.

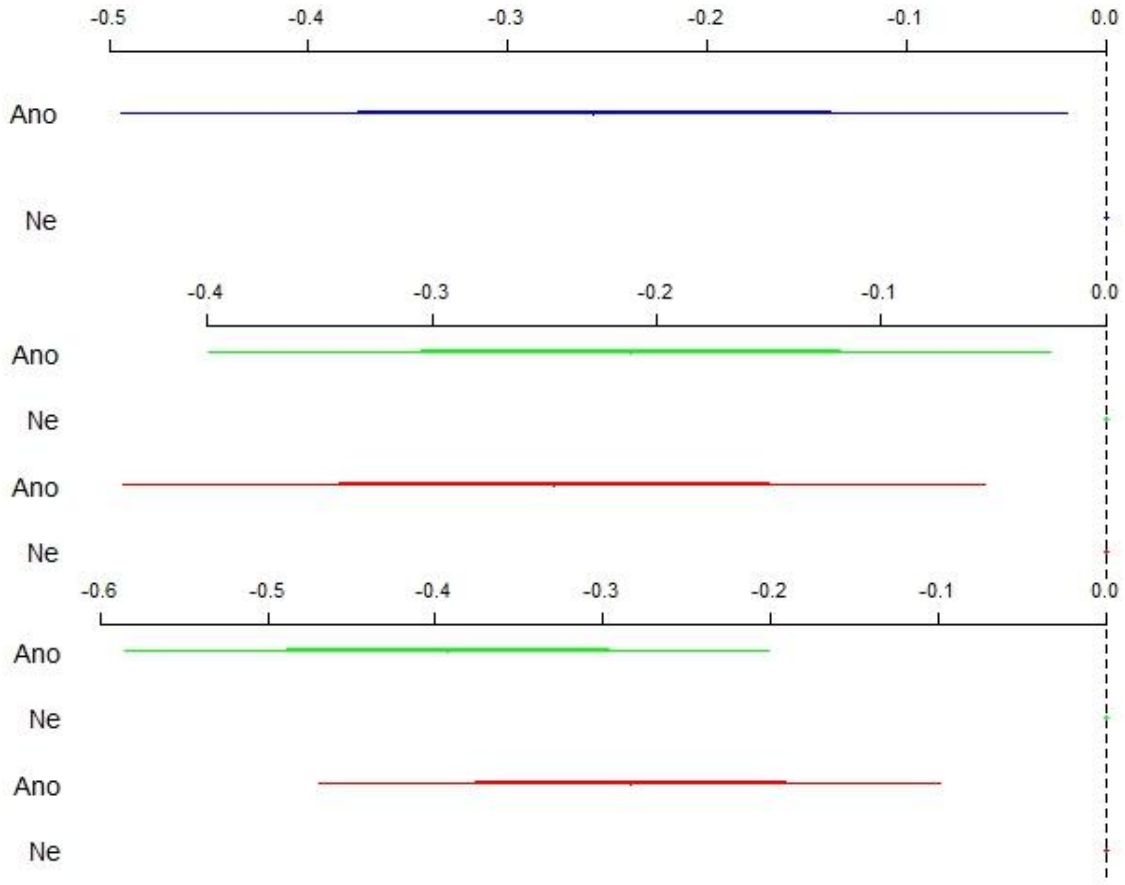
Představme si člověka, který je nezaměstnaný půl roku a teprve po půl roce se registruje na ÚP. V datech z VŠPS se počítá jeho doba nezaměstnanosti od začátku, zatímco ve výkazech z ÚP se počítá jeho doba od okamžiku registrace, tedy doba kratší o půl roku.

4.3.12 Pracovní zkušenost

V souboru 1 v období před krizí nacházeli dříve práci lidé, kteří hledali své první zaměstnání a to zhruba o 23 %. V dalších obdobích nebyl rozdíl statisticky významný. Základní výsledky jsou uvedeny v obrázku 4.28.

V souboru 2 byl statisticky významný rozdíl doby do nalezení práce mezi lidmi s předchozí pracovní zkušeností oproti lidem bez této zkušenosti v období během krize a po krizi. Během krize nacházeli práci dříve lidé bez pracovní zkušenosti zhruba o 22 % a po krizi o 19 %.

V souboru 3 byly výsledky obdobné jako v souboru 2 – lidé hledající první práci ji hledali během krize o zhruba 25 % kratší dobu a po krizi o 32 % kratší dobu.



Obrázek 4.28 Odhady parametrů podle existence pracovní zkušenosti

Lidé hledající první zaměstnání jej nachází dříve než lidé, kteří již pracovní zkušenost mají. Je velmi pravděpodobné, že tento výsledek bude souviset s věkem a vzděláním dotyčných, kdy první zaměstnání si hledají lidé obvykle po ukončení školy, tedy v mladém věku.

4.3.13 Typ hledaného úvazku

V souboru 1 trvalo nalezení práce lidem preferujícím spíše kratší typy úvazků déle než lidem preferujícím plný úvazek a to zhruba o 98 %. Během krize se rozdíly snížily, přičemž lidé hledající méně vyhraněné typy úvazků nacházeli práci zhruba o 29 % později než lidé preferující vyhraněné typy úvazků. V období po krizi se vyčlenila pouze kategorie lidí hledající spíše plný úvazek, ale akceptující i kratší – u nich byla doba do nalezení práce delší o zhruba 27 %. V období po krizi nicméně podle log-rank testu nevyšel rozdíl jako statisticky významný. Tyto výsledky jsou shrnuty v obrázku 4.29.

Z hlediska změn v průběhu času se statisticky významně změnila doba do nalezení práce pouze u lidí hledajících plný úvazek, u kterých narostla v období po krizi, viz tabulka A112 přílohy.

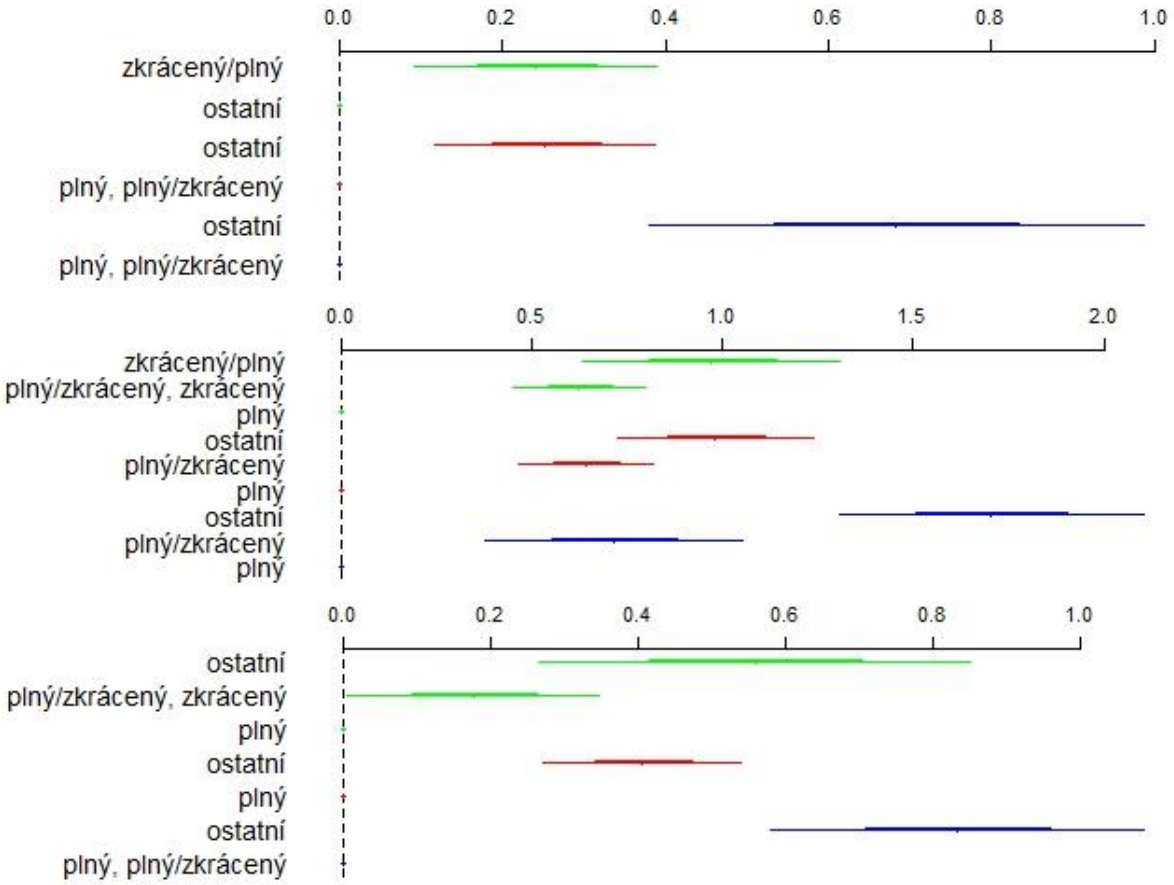
V souboru 2 v období před krizí hledali nejkratší dobu lidé hledající plný úvazek. Lidé akceptující i zkrácený úvazek hledali práci o zhruba 104 % déle, ostatní o 450 %. Během krize zůstalo rozdělení stejné, druhá skupina lidí hledala práci o 90 % déle a třetí o 167 % déle, rozdíly byly tedy menší. V období po krizi se skupiny lehce pozměnily, lidé se smíšenými preferencemi oproti lidem hledajícím pouze plný úvazek hledali o zhruba 87 % déle a lidé hledající jen zkrácený úvazek nebo jakýkoliv hledali o 164 % déle.

V žádné skupině nebyly v průběhu času zaznamenány statisticky významné pohyby v rámci různých období.

V souboru 3 v období před krizí měli statisticky významně nejkratší dobu nezaměstnanosti lidé preferující plný úvazek, lidé preferující kratší úvazky hledali práci o zhruba 130 % déle. Během krize nejkratší dobu nezaměstnanosti měli lidé hledající pouze plný úvazek, ostatní skupiny se od sebe nelišily, jejich doba nezaměstnanosti byla o 50 % delší. Po krizi byla stále nejkratší doba nezaměstnanosti u lidí hledajících pouze plný úvazek. Lidé hledající oba typy úvazků měli dobu nezaměstnanosti o zhruba 19 % vyšší, u lidí hledajících jen zkrácený úvazek byla doba nezaměstnanosti delší o 75 %.

V rámci různých období se hodnoty doby nezaměstnanosti statisticky významně lišily u lidí hledajících plný úvazek, u kterých poklesla doba nezaměstnanosti v období krize a pak

se zase vrátila na předkrizovou úroveň. Podobný vývoj zaznamenán i u skupiny lidí hledajících primárně zkrácený, ale akceptujících i plný úvazek, vše viz tabulka A120 přílohy.



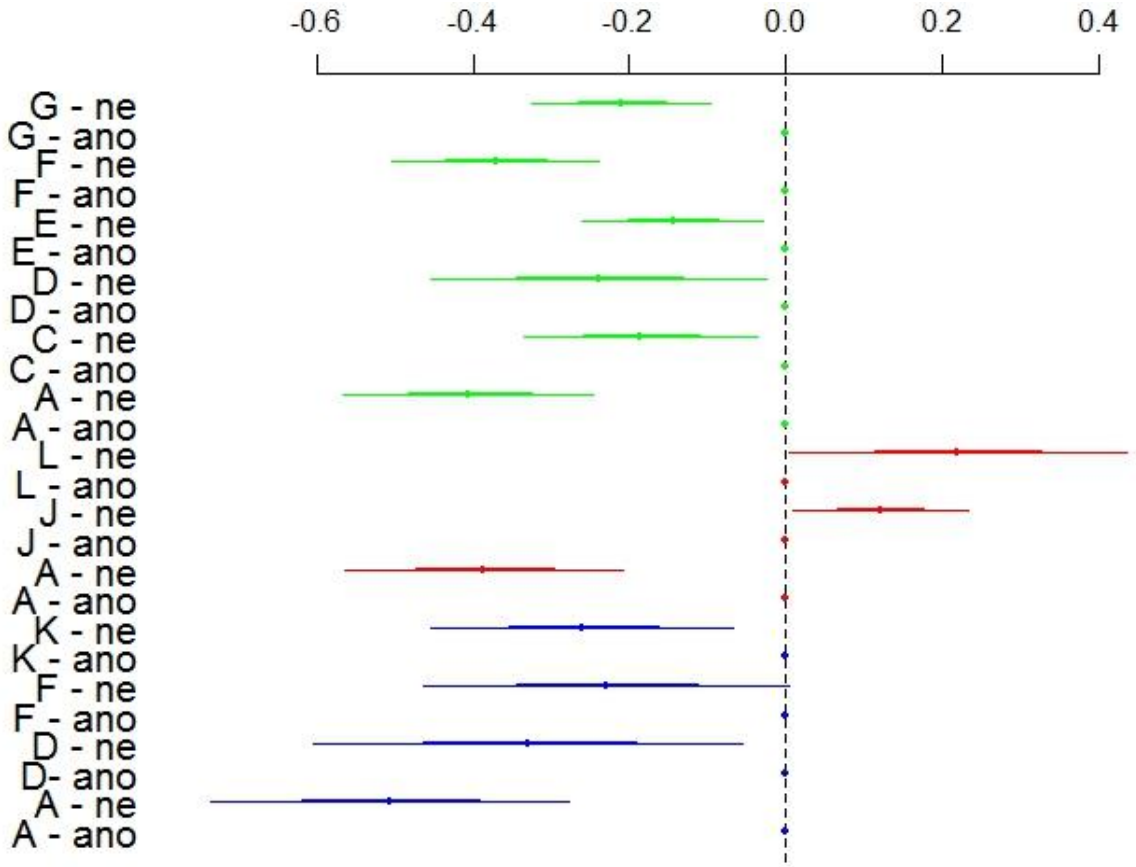
Obrázek 4.29 Odhady parametrů podle typu hledaného úvazku

Během krize došlo obecně ke zmenšení rozdílů v době do nalezení zaměstnání či doby nezaměstnanosti mezi jednotlivými typy preferovaných úvazků. V souboru 1 došlo k nárůstu doby do nalezení zaměstnání u lidí hledajících plný úvazek, což může být způsobeno reakcí firem na krizi.

4.3.14 Metody hledání

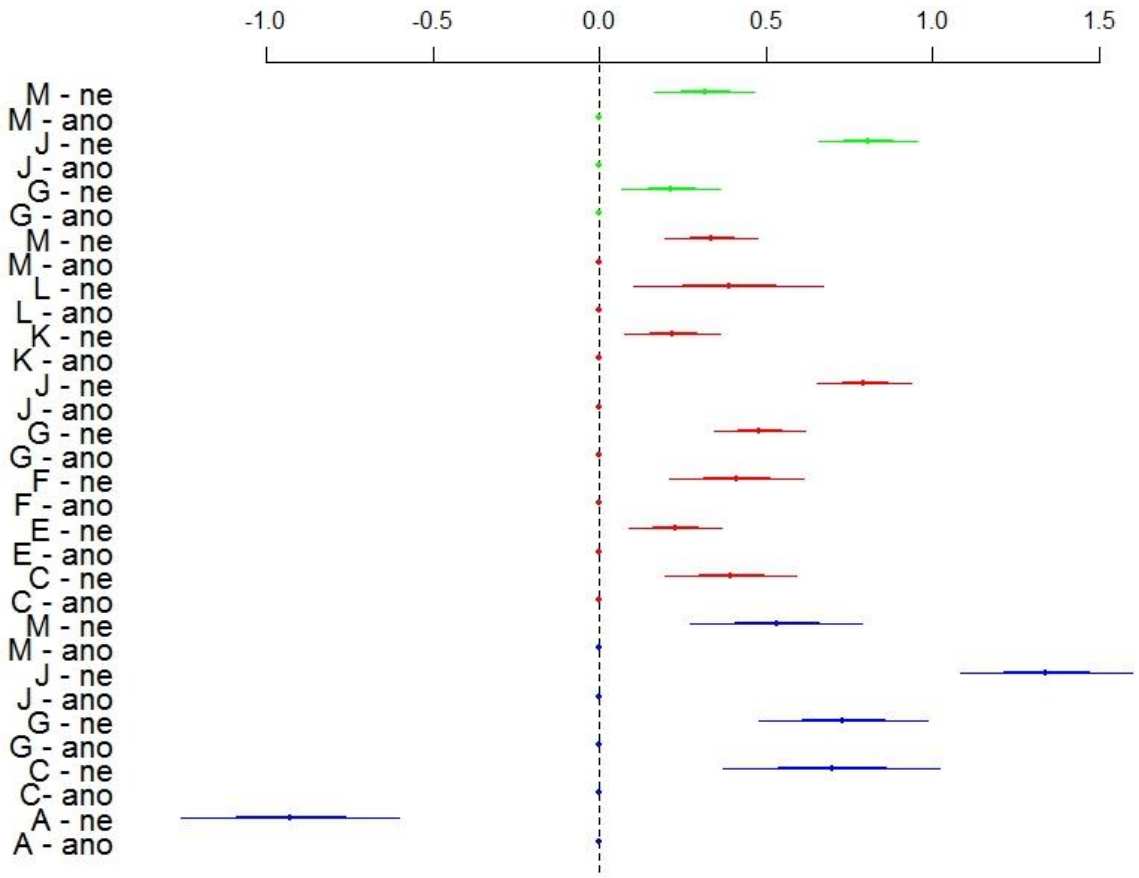
V období před krizí se statisticky významně lišila doba do nalezení zaměstnání v souboru 1, podle metod hledání A, D, F a K, ve kterých byla postupně kratší o 40 %, 28 %, 20 % a 23 % u těch, kteří danou metodu nepoužívali. Během krize byla tato doba kratší o 32 % u metody A a delší o 13 % u metody J a o 25 % u metody L. Po krizi byly již statisticky významné metody A, C, D, E, F, K – u všech byla doba do nalezení práce kratší u těch, kteří danou metodu nepoužívali a to postupně o 33 %, 17 %, 21 % 13 %, 31 % a 19 %. Výsledky za soubor 1 jsou souhrnně v obrázku 4.30.

Statisticky nevýznamně vždy vycházely rozdíly v době do nalezení práce podle metod B, I a M.



Obrázek 4.30 Odhady parametrů pro různé metody hledání práce, soubor 1

V souboru 2 byla před krizí doba do nalezení práce u lidí neužívajících metodu A kratší o zhruba 60 %. U ostatních metod byla naopak doba do nalezení práce delší, pokud člověk nepoužíval danou metodu – u metody C o 101 %, G o 108 %, J o 282 % a M o 70 %. V období krize byla statisticky významně delší doba do nalezení práce pro lidí nepoužívající danou metodu hned u 8 metod – u metody C o 48 %, u metody E o 25 %, u metody F o 51 %, u metody G o 62 %, u metody J o 121 %, u metody K o 25 %, u metody L o 47 % a metody M o 40 %. Po krizi byly naopak statisticky významné rozdíly jen u tří metod – G, J, M, vždy delší u těch, kteří danou metodu nepoužívali a to postupně o 24 %, 124 % a 37 %. Výsledky jsou souhrnně v obrázku 4.31.

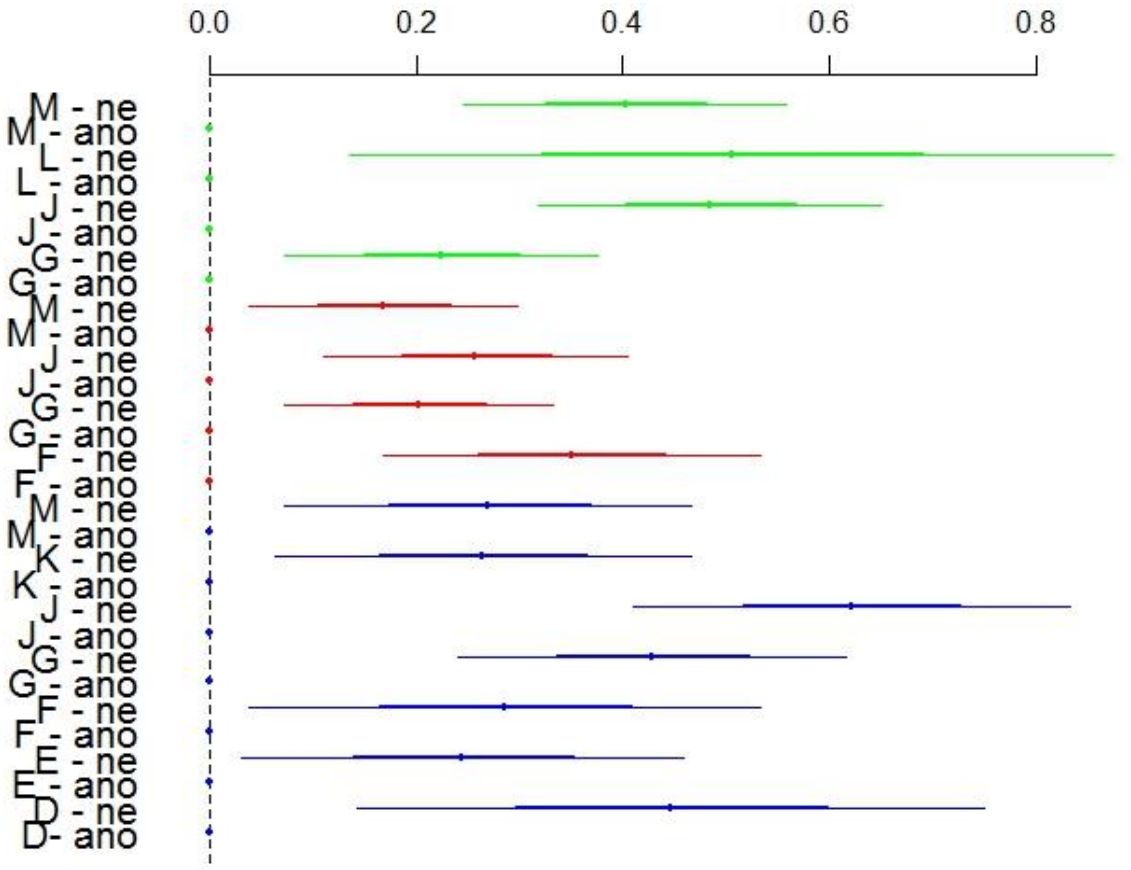


Obrázek 4.31 Odhady parametrů pro různé metody hledání práce, soubor 2

V souboru 3 byla doba nezaměstnanosti před krizí delší u těch, kteří nevyužívali metody hledání práce D, E, F, G, J, K a M a to postupně o 56 %, 28 %, 33 %, 53 %, 86 %, 30 % a 31 %. Během krize byly statisticky významné rozdíly u čtyř metod a to u F, G, J a M postupně 42 %, 22 %, 29 % a 18 % delší u těch, kteří tyto metody nevyužívali. Po krizi byla statisticky významně delší doba nezaměstnanosti u lidí, kteří nehledali práci pomocí čtyř metod – G, J, K a M a to 25 %, 62 %, 66 % a 50 %. Výsledky jsou v obrázku 4.32.

V prvním souboru měly všechny metody negativní vztah k době do nalezení zaměstnání, tedy jejich užití bylo spojeno s delší dobou. Tento vztah lze vysvětlit převrácenou kauzalitou, tedy že lidé hledající práci déle zkoušejí stále nové a nové metody. Výjimkou jsou metody J a L v období krize – obě dvě metody značí, že člověk očekává vyjádření potenciálního zaměstnavatele, tedy pasivní způsob hledání práce.

V dalších dvou souborech byla doba nezaměstnanosti naopak delší u lidí, kteří dané metody nepoužívali, s výjimkou metody A, což je hledání pomocí Úřadu práce – výsledky jsou konzistentní s registrací na ÚP.



Obrázek 4.32 Odhady parametrů pro různé metody hledání práce, soubor 3

4.3.15 Počet užitých metod hledání práce

V souboru 1 je vztah mezi dobou do nalezení práce a počtem použitých aktivních způsobů hledání práce kladný a lineární. V období během krize a po krizi je statisticky významný i vztah mezi dobou do nalezení práce a počtem způsobů jejího hledání užitého jako kategoriální proměnná, ale tento vztah není výrazně věrohodnější, aby ospravedlnil užití kategorizace.

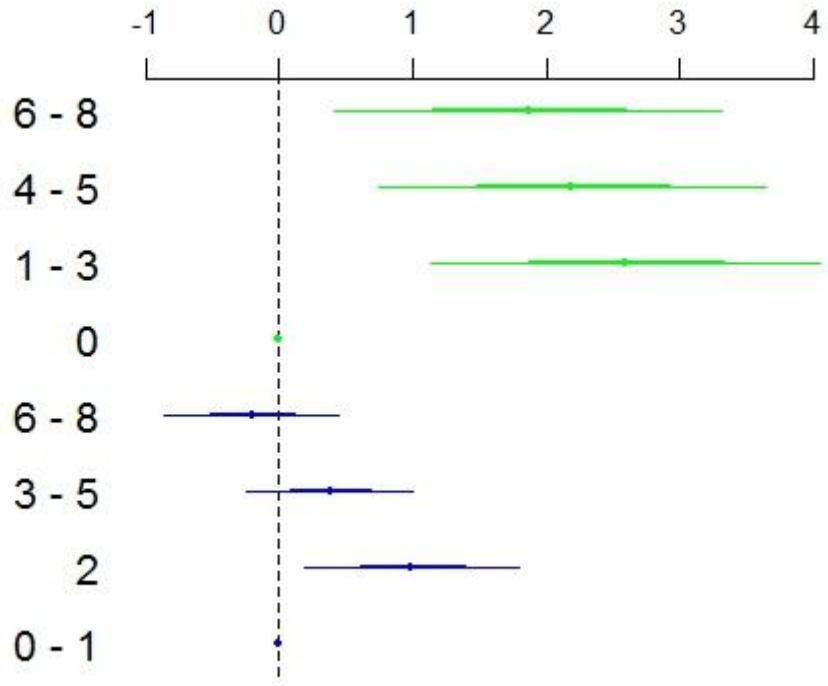
Před krizí byl každý způsob hledání práce navíc spojený s 6% prodloužením doby do nalezení práce a toto prodloužení bylo v dalších obdobích zhruba 3%, viz tabulka 4.28.

V souboru 2 byl směr tohoto vztahu složitější. Statisticky významný byl pouze v období před krizí a během krize. V období před krizí nacházeli práci nejrychleji lidé užívající 0, 1 anebo více než 5 způsobů hledání práce (rozdíl není statisticky významný). Nejpomaleji práci nacházeli lidé používající 2 způsoby – o 169 %. Lidé užívající 3 až 5 způsobů hledání práce nacházeli práci pomaleji o 47 %. V období krize nacházeli práci nejrychleji lidé nepoužívající ani jeden způsob hledání práce. Při 1 až 3 způsobech hledání práce byla doba do jejího nalezení

delší více než třináctinásobně. Při 4 nebo 5 pak skoro devítinásobně a při více metodách zhruba šesti a půlnásobně. Výsledky jsou v obrázku 4.33.

Tabulka 4.28 AFT model podle počtu užitých metod hledání práce, soubor 1

Hleda před krizí	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		4,21	0,58	14,89	2,16	7,82	2,11
Hleda	1,062	1,056						
Hleda během krize	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		4,98	0,44	14,37	1,34	7,56	1,75
Hleda	1,059	1,034						
Hleda po krizi	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		4,45	0,39	13,37	1,25	7,00	1,71
Hleda	1,080	1,032						



Obrázek 4.33 Odhady parametrů pro různé počty použitých metod hledání práce, soubor 2

Tabulka 4.29 AFT model podle počtu užitých metod hledání práce, soubor 3

Hleda před krizí	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		19,83	2,99	144,48	23,28	189,41	15,30
Hleda	0,865	1,065						
Hleda během krize	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		11,93	1,26	67,07	7,47	47,26	8,42
Hleda	0,943	1,042	11,25		63,22		44,55	7,94
Hleda po krizi	exp($x^T b$)	exp(s.e.*$u_{0,975}$)	$t_{0,5}$	s.e.	$t_{0,9}$	s.e.	E(T)	Modus
0	1,000		13,03	1,65	85,52	11,46	80,40	9,75
Hleda	0,938	1,050						

V souboru 3 nebyl statisticky významný rozdíl věrohodnosti mezi modely užívajícími počet způsobů hledání práce jako kategoriální proměnnou a jako kvantitativní proměnnou. Zde vychází, že čím více se používá způsobů hledání práce, tím kratší je doba nezaměstnanosti. V období před krizí je přidání jednoho způsobu hledání práce spojeno s dobou nezaměstnanosti kratší o zhruba 13 %, v období krize a po krizi je za každý způsob hledání práce doba nezaměstnanosti nižší o 6 %. Výsledky jsou v tabulce 4.29.

Výsledky jsou konzistentní se zkoumáním jednotlivých metod hledání práce – v souboru 1 platí, že čím více metod je používáno, tím delší je doba do nalezení práce, což vysvětluji tezí o obrácené kauzalitě. V souboru 2 je použito kategoriální členění a výsledky ukazují, že čím více je použito metod, tím kratší je doba do nalezení práce s výjimkou těch, kteří nepoužili žádnou metodu – tady opět můžeme uvažovat o tom, že takoví lidé ještě nehledají práci příliš dlouho. Pro tento rozpor nemám žádné vysvětlení. V souboru 3 jsou výsledky obdobné.

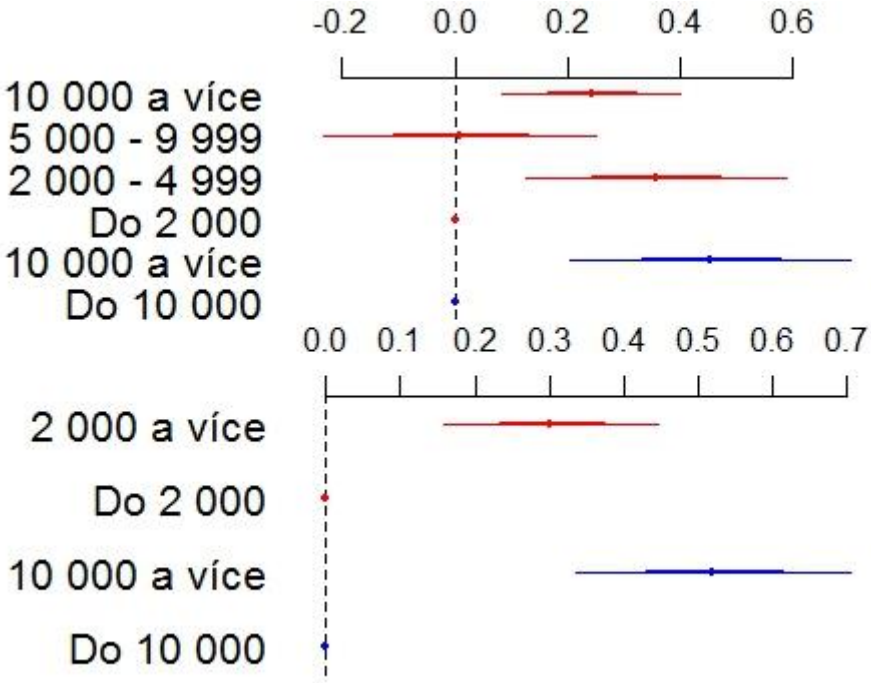
4.3.16 Velikost obce

V souboru 1 nebyl nalezen statisticky významný vztah k počtu obyvatel místa bydliště v žádném ze tří sledovaných období.

V souboru 2 v období před krizí nacházeli dříve práci lidé žijící v obcích do 10 000, v obcích nad 10 000 obyvatel byla doba do nalezení práce delší o zhruba 58 %. V období krize byla nejkratší doba do nalezení práce u lidí žijících v obcích do 2 000 obyvatel a v obcích

s 5 000 až 9 999. U lidí žijících v obcích s počtem obyvatel 10 000 a více byla doba do nalezení práce vyšší zhruba o 28 % a u lidí v obcích se 2 000 a 4 999 obyvateli vyšší o cca 43 %. Výsledky jsou prezentovány v obrázku 4.34.

V souboru 3 byla v období před krizí doba nezaměstnanosti kratší u lidí žijících v obcích do 9 999 obyvatel, ve větších obcích byla delší o 68 %. Během krize se pak dala statisticky významně odlišit doba nezaměstnanosti u lidí v obcích do 1 999 obyvatel a nad tento počet, kdy lidé ve větších městech měli dobu nezaměstnanosti delší o 35 %. Výsledky jsou v obrázku 4.34.



Obrázek 4.34 Odhady parametrů pro různé počty obyvatel

Výsledky obecně hovoří o tom, že lidé z větších měst mají delší dobu do nalezení zaměstnání nebo dobu nezaměstnanosti. To může být způsobeno korelací velikosti sídla s jeho polohou v určitém regionu – například v regionech Severozápad a Moravskoslezsko s obecně delší dobou nezaměstnanosti zahrnují poměrně velký počet větších měst.

Každopádně výsledky nesvědčí ve prospěch intuitivní hypotézy, že obyvatelé větších měst nachází práci rychleji.

4.4 Minimální adekvátní model

Výsledky minimálních adekvátních modelů (MAM) pro souboru 1, 2 a 3 umožňují vidět závislosti i při zahrnutí dalších vlivů. Počáteční nastavení základních kategorií modelů je

shrnuto v tabulce 4.30. Tyto kategorie byly zvoleny podle jejich četností, tedy za základní kategorii byla vybrána ta nejčetnější. Výjimkou je *Období*, kde bylo zvoleno období před krizí.

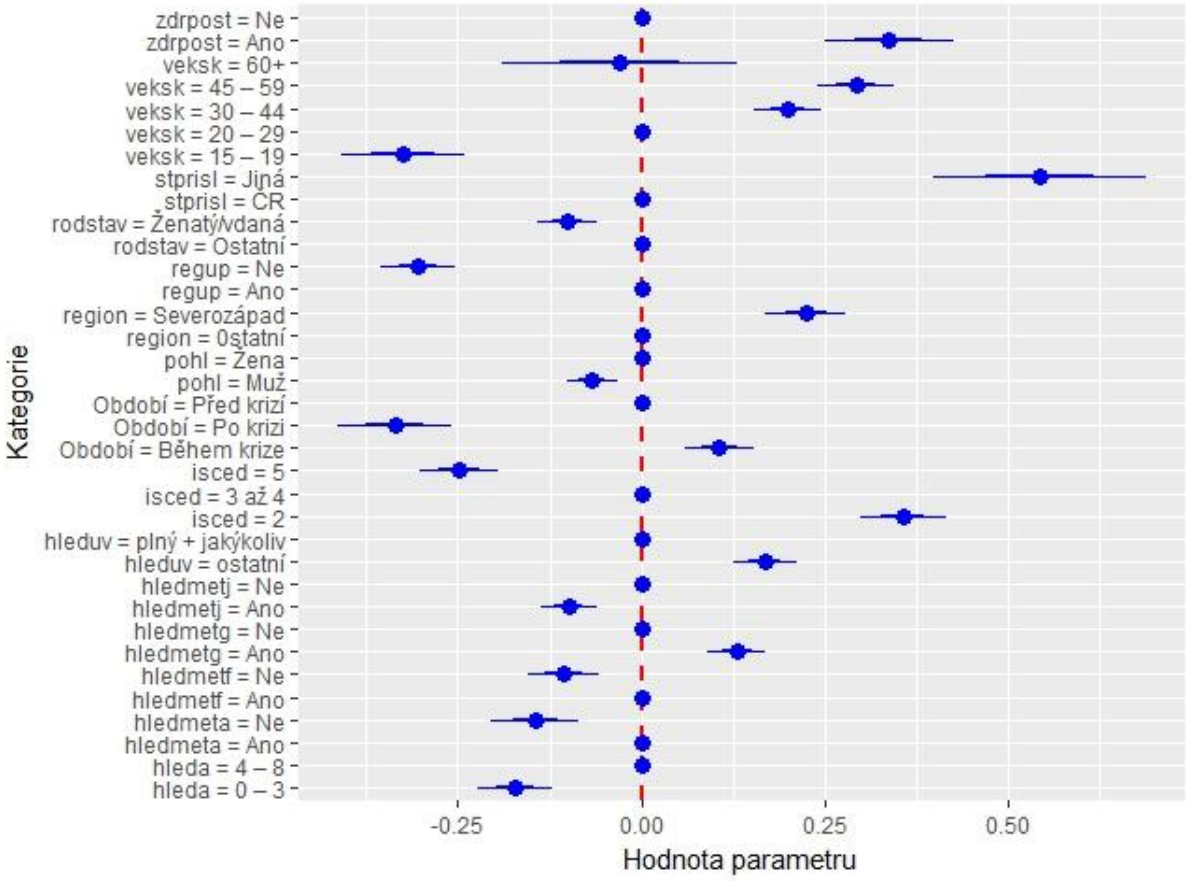
V obrázku 4.35 je potom zachycen samotný minimální adekvátní model pro soubor 1, v obrázcích 4.36 a 4.37 jsou tyto modely pro soubor 2 a 3.

Tabulka 4.30 Základní kategorie

Proměnná	Kategorie	Proměnná	Kategorie	Proměnná	Kategorie	Proměnná	Kategorie
Období	Před krizí	zemnar	ČR	hledmeta	Ano	hledmetj	Ne
Region	Jihovýchod	stprisl	ČR	hledmetb	Ne	hledmetk	Ne
Pocod	3	zdrpost	ne	hledmetc	Ano	hledmetl	Ne
Vzocd	Osoba v čele domásnosti	isced	3	hledmetd	Ano	hledmetm	Ne
pohl	Žena	regup	Ano	hledmete	Ne	hleda	6
rodstav	Svobodný/á	exzam	Ano	hledmetf	Ano	velikostobce	10 a více
veksk	20 – 24	hleduv	plná doba	hledmetg	Ne		

- V souboru 1 udává MAM rozdíly v rámci jednotlivých období – v období během krize byla doba do nalezení zaměstnání delší o 11 %, v období po krizi naopak o 28 % kratší. Tyto výsledky jsou v rozporu s výsledky v části 4.1, což může být způsobeno vlivem struktury.
- Ženatí či vdané nacházejí práci o 10 % dříve než ostatní, což je opět v rozporu s předchozími výsledky. V MAM nejspíš došlo k započítání vlivu věku, což výsledek obrátilo.
- Jako výrazný faktor, dokonce nejvýraznější ze všech, se objevuje státní příslušnost. Lidé ze zahraničí hledají práci déle o více než 70 %.
- Nejdéle trvá nalézt práci v Severozápadním regionu.
- Muži nalézají práci o něco dříve než ženy, ale rozdíl je na hranici hladiny významnosti.
- Nejrychleji nachází práci mladí, s rostoucím věkem se doba do nalezení práce prodlužuje.
- Zdravotně postižení nacházejí práci později, podobně lidi s nižším vzděláním, přičemž nelze odlišit dobu do nalezení práce u lidí s maturitou a bez maturity.

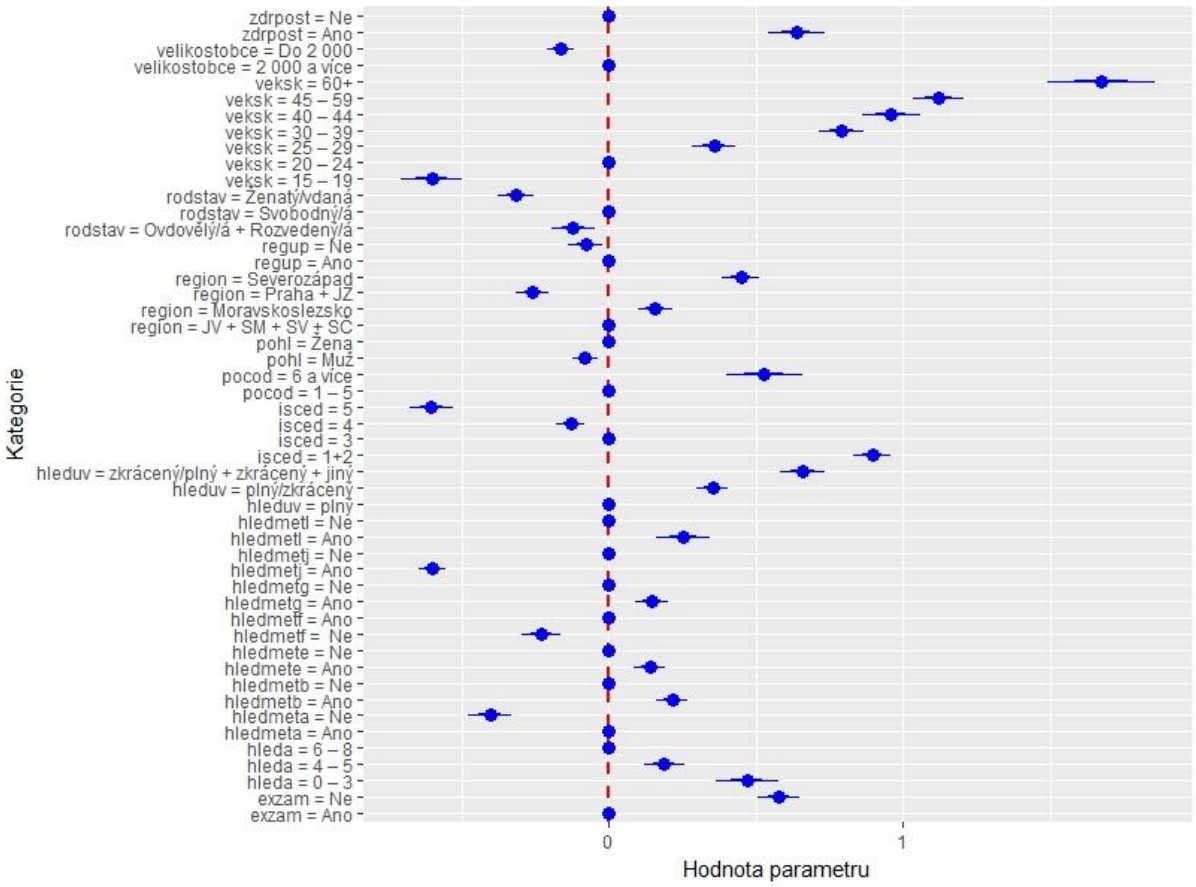
- Lidi registrovaní na ÚP nachází práci později, podobně lidi hledající práci přes ÚP a inzeráty.
- Naopak lidé chodící k pohovorům nachází práci o něco dříve.
- Větší množství metod použitých k hledání práce je spojeno s delší dobou nezaměstnanosti, což může být zdůvodněno tím, že lidé déle hledající práci zkouší více způsobů. To by znamenalo, že na počátku nezaměstnanosti zkouší člověk najít práci menším množstvím metod než později, kdy už je nezaměstnaný déle.



Obrázek 4.35 Odhady parametrů minimálního adekvátního modelu, soubor 1

- V souboru 2 nebyl v MAM nalezen statisticky významný rozdíl v době do nalezení práce podle období. To může být způsobeno faktem, že nově příchozí nezaměstnaní snižují dobu nezaměstnanosti v tomto souboru, což je na druhé straně vyváženo obecně zhoršenou uplatnitelností na trhu práce v tomto období.
- Svobodní nalézali práci nejpomaleji ze všech, což je v rozporu s výsledky individuálního zkoumání v kapitole 4.3. Opět zde zřejmě hraje roli zahrnutí vlivu věku.

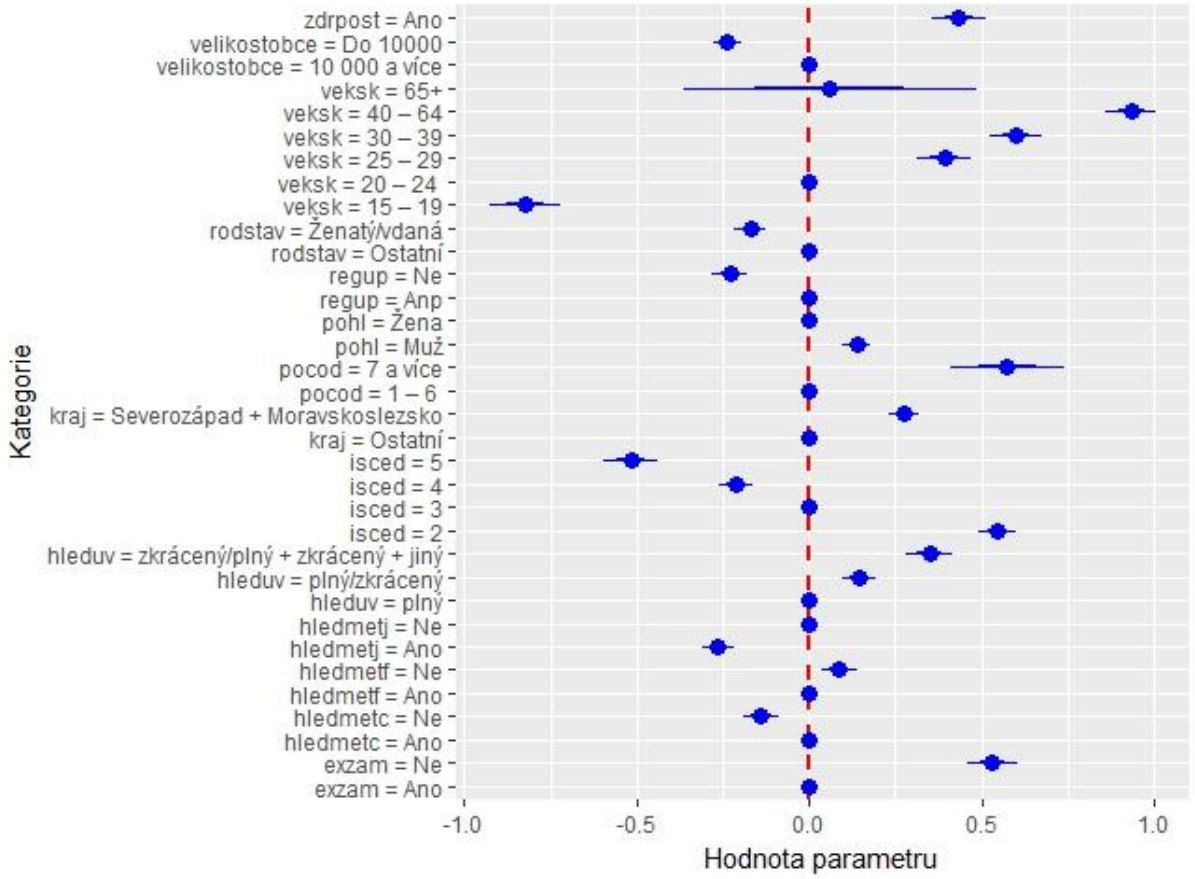
- Registrace na Úřadu práce je spojena s delší dobou nezaměstnanosti, podobně jako v souboru 1, ale méně výrazně.
- Další výsledky v rozporu jsou hodnoty koeficientů pro jednotlivé způsoby hledání práce. S delším hledáním práce je spojeno hledání práce přes ÚP, soukromé zprostředkovatele, podávání a sledování inzerátů, účast na pohovorech. Naopak očekávání výsledku přijímacího řízení je spojeno s výrazně kratší dobou nezaměstnanosti – o 45 %.
- Naopak platí, že čím méně se použije způsobů hledání práce, tím delší je doba do nalezení práce.
- Výrazným činitelem je vzdělání – čím vyšší, tím kratší doba do nalezení zaměstnání.
- Nejpomaleji se práce nachází v regionu Severozápad, poté Moravskoslezsko, nejrychleji naopak v Praze a regionu Jihovýchod.
- V domácnostech se 6 a více členy je doba do nalezení práce asi o 70 % delší.
- Čím starší je nezaměstnaný, tím pomaleji si nachází práci. Zde nedochází k anomálii v nejstarší skupině.
- Zdravotně postižení hledají práci zhruba o 89 % déle.
- V modelu existuje i výrazný vliv pracovní zkušenosti v nesouladu s předchozím zjištěním, tedy že lidé hledající první práci ji dříve nachází. V MAM je to naopak, což je zřejmě dáno zahrnutím vlivu věku do modelu.
- Obyvatelé nejmenších obcích do 2 000 obyvatel nachází zaměstnání nejdříve.
- Čím delší je hledaný úvazek, tím kratší je doba nezaměstnanosti.



Obrázek 4.36 Odhady parametrů minimálního adekvátního modelu, soubor 2

- V souboru 3 stejně jako v souboru 2 není významný vliv období, tedy není rozdíl v době nezaměstnanosti před, během a po krizi.
- Nejdéle trvá nalézt práci obyvatelům regionů Severozápad a Moravskoslezsko.
- Lidé z domácností se 7 a více členy hledají práci nejdéle.
- V tomto souboru jsou muži déle nezaměstnaní než ženy. To jenom podtrhuje celkově nevýrazné rozdíly v době nezaměstnanosti mezi pohlavími.
- Za jinak stejných okolností nacházeli ženatí práci o něco dříve.
- Starší lidé obecně mají vyšší dobu nezaměstnanosti s výjimkou nejstarší kategorie.
- Vyšší vzdělání je statisticky významně spojeno s kratší dobou nezaměstnanosti.
- Registrace na ÚP je spojena s delší dobou nezaměstnanosti. Tento výsledek je v rozporu s individuálním zkoumáním, rozdíl může být způsobený strukturou.
- Preference plného úvazku je spojena s kratší dobou nezaměstnanosti.

- Zkušenost s předchozím zaměstnáním je spojena s výrazně kratší dobou nezaměstnanosti – rozdíl proti předchozímu zjištění je zřejmě dán zahrnutím věkové struktury.
- Lidé z menších obcí (Do 10 000 obyvatel) jsou nezaměstnaní kratší dobu.
- Z hlediska způsobů hledání práce je přímé kontaktování zaměstnavatelů spojeno s delší dobou nezaměstnanosti. Naopak s kratší dobou nezaměstnanosti je spojeno čekání na výsledky pracovního řízení.



Obrázek 4.37 Odhady parametrů minimálního adekvátního modelu, soubor 3

4.5 Modely s interakcemi

V této kapitole se zabývám interakcemi zahrnutými do minimálních adekvátních modelů z kapitoly 4.4 postupem uvedeným v kapitole 3.8. Všechny interakce popisuji v jednotlivých bodech. Obrázky příslušené popisovaným efektům jsou v Příloze na konci disertační práce. Je důležité mít na paměti, že některé z výsledků můžou být pouze náhodnými artefakty metody

odvozování regresního modelu. Plné specifikace těchto modelů jsou k dispozici v elektronické příloze C.

V souboru 1 jsem identifikoval celkem 6 interakcí významných na zvolené hladině významnosti 0,01. Tyto interakce jsou

1. Období a region. Tato interakce potvrzuje již dříve uvedené tvrzení o sbližování doby do nalezení práce v jednotlivých regionech. V regionu Severozápad se doba nezaměstnanosti průběžně snižovala, zatímco v ostatních regionech vzrostla v období krize a poté poklesla. (Obrázek 1 Přílohy)
2. Pohlaví a věk. U žen doba do nalezení zaměstnání s rostoucím věkem roste výrazně více než u mužů. (Obrázek 2 Přílohy)
3. Registrace na Úřadu práce a hledání metodou J (osoba očekává výsledek vyřízení žádosti o práci). U lidí registrovaných na Úřadu práce nebylo čekání na výsledek vyřízení žádosti o práci statisticky významným faktorem, zatímco u neregistrovaných osob ano. (Obrázek 3 Příloha)
4. Rodinný stav a hledání metodou F (osoba hledá práci prohlížením inzerátů). Ženatí či vdané mají významně nižší dobu do nalezení práce, pokud nehledají metodou F, oproti ostatním možným kombinacím. (Obrázek 4 Přílohy)
5. Věk a vzdělání. Ve vyšším věku jsou nižší rozdíly mezi různými úrovněmi vzdělání. (Obrázek 5 Přílohy)
6. Věk a hledání metodou A (pomocí ÚP). Mladší osoby nachází práci dříve, pokud nehledají práci pomocí ÚP, u starších osob rozdíly nejsou pozorovatelné. (Obrázek 6 Přílohy)

V souboru 2 jsem identifikoval celkem 11 statisticky významných interakcí.

1. Region a vzdělání. Vzdělání hrálo nejmenší roli z hlediska doby do nalezení práce v regionech Praha a Jihozápad, naopak nejvýraznější v regionech Moravskoslezsko a Severozápad. (Obrázek 7 Přílohy)
2. Region a hledání metodou E (podávání inzerátů). Podávání inzerátů působilo pozitivně (bylo spojeno s nižší dobou do nalezení zaměstnání) v regionech Praha, Jihozápad, Moravskoslezsko a Severozápad, naopak negativně v ostatních regionech. (Obrázek 8 Přílohy)

3. Rodinný stav a vzdělání. Vzdělání hraje relativně menší roli u svobodných než u ostatních skupin obyvatel. Tento výsledek byl dosažen i přes to, že u mladších (a tedy spíše svobodných) lidí hraje vzdělání vyšší roli. (Obrázek 9 Přílohy)
4. Věk a vzdělání. Role vzdělání z hlediska doby do nalezení práce klesá s rostoucím věkem. Toto je v souladu s výsledkem v souboru 1. (Obrázek 10 Přílohy)
5. Vzdělání a typ úvazku. U lidí hledajících kratší úvazky hrálo vzdělání menší roli. (Obrázek 11 Přílohy)
6. Hledání metodou A a F (pomocí úřadu práce a odpovídání na inzeráty). Lidi nehledající práci odpovídáním na inzeráty nacházeli práci dříve, pokud zároveň nehledali práci s pomocí Úřadu práce. To je v souladu s ostatními zjištěními, že hledání práce pomocí ÚP je spojeno s delší dobou nezaměstnanosti. Odpovídání na inzeráty je opět spojen s delší dobou do nalezení práce, jedná se nejspíše o indikátor doby hledání práce, tedy o metodu používanou až při déle trvající nezaměstnanosti. (Obrázek 12 Přílohy)
7. Věk a registrace na Úřadu práce. U starších lidí je registrace na Úřadu práce spíše spojeno s kratší dobou do nalezení zaměstnání než u mladších lidí. (Obrázek 13 Přílohy)
8. Hledání metodou F a vzdělání. U lidí odpovídajících na inzeráty hraje vzdělání o něco vyšší roli z hlediska doby do nalezení práce. (Obrázek 14 Přílohy)
9. Hledání metodou A a J. U lidí hledajících zaměstnání pomocí Úřadu práce nebylo čekání na výsledek vyřízení žádosti o práci tak významným faktorem jako u těch, co Úřad práce nepoužívali. (Obrázek 15 Přílohy)
10. Hledání metodou B (pomocí soukromých zprostředkovatelů) a G (pomocí pohovorů). Ti, kteří používali obě metody zároveň, měli dobu do nalezení práce významně vyšší než ostatní skupiny. Jedná se zde možná o indikátor a ne o kauzální faktor. (Obrázek 16 Přílohy)
11. Registrace na ÚP a vzdělání. U lidí registrovaných na Úřadu práce hrálo vzdělání menší roli než u neregistrovaných. (Obrázek 17 Přílohy)

V souboru 3 jsem identifikoval celkem 7 statisticky významných interakcí.

1. Věk a vzdělání. Opět to vypadá, že věk hraje u starších lidí menší roli. Výjimkou můžou být lidé důchodového věku. (Obrázek 18 Přílohy)
2. Pracovní zkušenost a vzdělání. U lidí bez pracovní zkušenosti hraje vzdělání větší roli než u lidí s pracovní zkušeností. (Obrázek 19 Přílohy)

3. Velikost obce a vzdělání. U větších obcí se zdá, že vzdělání hraje větší roli než u menších. (Obrázek 20 Přílohy)
4. Velikost obce a hledání metodou C (přímé kontaktování zaměstnavatele). Lidé v menších obcích kontaktující zaměstnavatele přímo mají delší dobou nezaměstnanosti. Ve větších obcích nehraje tato proměnná roli. V tomto případě se opět lze domnívat, že přímé kontaktování je známkou toho, že dotyčný zkouší metody, které na začátku nezaměstnanosti nezkoušel. (Obrázek 21 Přílohy)
5. Registrace na Úřadu práce a hledání metodou J (osoba očekává výsledek vyřízení žádosti o práci). U lidí registrovaných na Úřadu práce nebylo čekání na výsledek vyřízení žádosti o práci tak významným faktorem jako u neregistrovaných osob. (Obrázek 22 Přílohy)
6. Hledání metodou F a vzdělání. U lidí odpovídajících na inzeráty hraje vzdělání menší roli. (Obrázek 23 Přílohy)
7. Registrace na ÚP a vzdělání. U lidí registrovaných na ÚP hraje vzdělání menší roli z hlediska doby nezaměstnanosti. (Obrázek 24 Přílohy)

Celkově lze ve všech třech datových souborech najít tři shodné rysy:

- Vzdělání hraje roli především u mladších lidí.
- U lidí registrovaných na Úřadu práce či s jeho pomocí hledajících práci bylo očekávání výsledku žádosti o práci méně významným faktorem z hlediska doby do nalezení práce. Vysvětlením může být, že lidé hledající práci pomocí ÚP jsou častěji odmítáni.
- Občas se vyskytují těžko vysvětlitelné kombinace vykazující statistickou významnosti. Důvodem může být např. fakt, že v daném modelu jsou vztahy platné pro nulové hodnoty ostatních proměnných, tedy fakticky za předpokladu, že hodnoty ostatních proměnných jsou nastaveny na referenční (viz tabulka 4.30 pro počáteční referenční kategorie). Poté se tedy může jednat o pouhou náhodu danou velkým množstvím možností.

Závěr

V disertační práci jsem se zabýval využitím analýzy přežití pro zhodnocení doby do nalezení zaměstnání, respektive doby nezaměstnanosti.

Data pro práci pocházejí z Výběrového šetření pracovních sil, přičemž doba nezaměstnanosti v těchto datech je intervalově cenzorovaná náhodná veličina. Proto jsem využil metody analýzy přežití specifické pro tento typ cenzorování, jmenovitě z neparametrických metod Turnbullův odhad a vážený log-rank test podle Suna, z parametrických metod potom regresní AFT model s využitím log-logistického rozdělení.

Kromě těchto metod jsem vylepšil odhad pravých konců rozdělení doby nezaměstnanosti využitím metody špiček nad prahem mající původ v teorii extrémní hodnoty v souboru obsahujícím pouze ty, kteří nalezli zaměstnání.

Většina výše zmíněných metod není v české literatuře obvykle popsána dostatečně pro účely této práce, proto jsem věnoval pozornost i podrobnému zavedení těchto metod.

Pro účely práce jsem z dostupných dat vytvořil tři soubory pozorování, přičemž první z nich obsahuje pouze ty lidi, kteří nalezli ve sledovaném období zaměstnání, druhý obsahuje všechny nezaměstnané ve sledovaném období bez ohledu na to, jestli práci nalezli nebo ne, a třetí obsahuje všechny pozorování ve vybraných čtvrtletích. Tato volba tři odlišných datových souborů umožňuje komplexnější zhodnocení celé problematiky a zhodnocení toho, co který přístup umožňuje a v čem se liší výsledky.

V práci pracuji se třemi obdobími, nazvanými jako období *před krizí*, *během krize* a *po krizi*. Díky tomuto rozlišení mi práce umožnila zkoumat vliv poslední ekonomické krize na dobu do nalezení práce a dobu nezaměstnanosti.

Z hlediska splnění cílů stanovených v úvodu konstatuji, že cíle byly splněny:

- Aplikoval jsem metody analýzy přežití na data z Výběrového šetření pracovních sil – konkrétní metody byly zmíněny výše.
- Navrhnul jsem metodiku pro tvorbu modelu doby do nalezení zaměstnání a doby nezaměstnanosti v závislosti na dostupných faktorech – popsal jsem tvorbu a vytvořil jsem úplný model a minimální adekvátní model pro dobu do nalezení zaměstnání a dobu nezaměstnanosti.
- Popsal jsem vývoj pravděpodobnostního rozdělení doby nezaměstnanosti v čase s ohledem na ekonomickou situaci pomocí AFT modelu a tento srovnal s modelováním jednotlivých období samostatně.

- Odhadnul jsem pravé konce pravděpodobnostních rozdělení doby do nalezení práce pomocí aplikace teorie extrémní hodnoty na intervalově cenzorovaná data.

Z hlediska samotných výsledků zejména z minimálních adekvátních modelů mohu konstatovat určité zákonitosti, které se projevují napříč všemi třemi datovými soubory.

- Obecně nachází o něco dříve práci ženatí či vdané.
- Nalézt práci trvá nejdéle v regionech Severozápad a Moravskoslezsko.
- Rozdíly mezi pohlavími jsou relativně vůči jiným vlivům malé, pohybují se okolo 10 %.
- S rostoucím věkem roste i doba do nalezení práce či doba nezaměstnanosti.
- S rostoucím vzděláním naopak tato doba klesá, přičemž vzdělání je velmi výrazný faktor.
- Zdravotně postižení lidé hledají práci déle.
- Lidé z domácností s větším počtem členů, obvykle více než 5, hledají práci déle – tuto proměnnou nejspíš můžeme považovat za určitý indikátor sociálního postavení.
- Registrace na ÚP je spojena s delší dobou nezaměstnanosti. To se zdá být s rozporu se zjištěními v kapitole 1.3, ale možnou příčinou tohoto nesouladu je zpoždění v hlášení se na úřadu práce.

Některé výsledky se pak vyskytují jenom v určitých souborech nebo jsou dokonce v rozporu:

- Vliv krize je zaznamenatelný pouze u těch, kteří práci nalezli. V souborech se všemi nezaměstnanými naopak krize vliv na pravděpodobnostní rozdělení doby nezaměstnanosti neměla. To může být způsobeno příchodem nových nezaměstnaných, kteří snižují průměrnou dobu nezaměstnanosti a další charakteristiky.
- Počet způsobů hledání práce je v prvním souboru negativně spojen s dobou do nalezení zaměstnání, v dalších dvou souborech je to opačně.
- Obyvatelé menších obcí nalézají práci o něco dříve.
- Čím delší je preferovaný úvazek, tím rychleji je práce nalezena.
- Zkušenost s předchozím zaměstnáním je spojena s rychlejším nalezením nového.

- V souboru 1 se jako nejvýraznější faktor vyskytuje státní příslušnost.

Interpretace jednotlivých zjištění je nicméně složitá, neboť doba nezaměstnanosti tak, jak je pojata v souborech 2 a 3 je silně ovlivněna příchozími nezaměstnanými. Tedy jestliže například nenacházím vliv státní příslušnosti na nezaměstnanost v těchto dvou souborech, může to být způsobeno tím, že bylo propuštěno relativně více cizích státních příslušníků než Čechů a průměrná doba nezaměstnanosti v této skupině tak klesá rychleji než u Čechů. Soubor 1 tímto interpretačním problémem netrpí.

Tato práce umožňuje aplikovat zde prezentované metody a tvorbu modelu jak na další časové úseky v rámci národních dat, tak na mezinárodní úrovni v závislosti na dostupnosti dat z jednotlivých národních LFS (*Labour Force Survey*) a provést tak další srovnání.

Seznam literatury a zdrojů

1. AALEN, O. (1978). Nonparametric Inference for a Family of Counting Processes. *Annals of Statistics*, 6.
2. AKAIKE, H (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory*. Tsahkadsor, Arménie (SSSR). 2. – 8. 9. 1971. Budapešť.
3. AKAIKE, H. (1974). A New Look at the Statistical Model Identifiaction. *IEEE Translations of Automatic Control*. 19 (6), 716 – 723. doi:10.1109/TAC.1974.1100705
4. ALTMAN, D.G., ANDERSEN, P.K. (1986). Bootstrap Investigation of the Stability of a Cox Regression Model. *Statistics in Medicine*. 8.
5. ANDERTON, A. (2006) *Economics*. 4th ed. Ormskirk: Causeway. ISBN 19-027-9692-6.
6. BENDEL R.B., AFIFI, A.A. (1977). Comparison of Stopping Rules in Forward Regression. *Journal of American Statistical Association*. 72.
7. BETENSKY, R. LINDSEY, J., RYAN, L, WAND, M, (2002). A Local Likelihood Proportional Hazards Model for Interval Censored Data. *Statistics in Medicine*. 21.
8. BLANCHARD, O. – PORTUGAL, P. (2001): What Hides Behind an Unemployment Rate: Comparing Portugese and U. S. Labor Markets. *American Economic Review*, 91, č. 1, str.. 187–207
9. BRICK, K. MLATSBENI, C. (2008). Examining the Degree of Duration Dependence in the Cape Town Labour Market. *A Southern Africa Labour and Development Research Unit Working Paper*, 10, 1-38. Cape Town: SALDRU, University of Cape Town.
10. CAI, T. BETENSKY, R. (2003). Hazard Regression for Interval-censored Data with Penalized Spline. *Biometrics*, 59.
11. CASTILLO, E. HADI, A. (1997). Fitting the Generalized Pareto Distribution to Data. *Journal of the American Statistical Association, 92*(440), 1609-1620.
12. CAZES, S. & SCARPETTA, S. (1998). Labour market transitions and unernployment duration: Evidence from Bulgarian and Polish micro-data. *Economics of Transition*, 6(I), 113- 144.
13. CHATFIELD, C. (1995). Model Uncertainty, Data Mining and Statistical Inference (with Discussion). *Journal of the Royal Statistical Society A*. 158.

14. CHEN, D., SUN, J., PEACE, K.E. (2013) *Interval-censored time-to-event data: methods and applications*. Boca Raton, FL: CRC Press. ISBN 978-146-6504-257.
15. COPAS, J.B. LONG, T. Estimating the Residual Variance in Orthogonal Regression with Variable Selection. *Journal of the Royal Statistical Society B*. 20.
16. COSTANZA M.C, AFIFI, A.A. (1979). Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis. *Journal of American Statistical Association*. 74.
17. COX, D.R. (1972) Regression Models and Life Tables (with Discussion). *Journal of Royal Statistical Society,* B30.
18. CUESTA T., & GONZÁLES, M. (2014). Análisis de los determinantes del desempleo y su duración en el Ecuador, periodo 2007-2012. *Escuela Politécnica Nacional*.
19. ČABLA, A. (2010) *Odhady v analýze přežívání*. Praha. Diplomová práce. Vysoká škola ekonomická v Praze. Vedoucí práce Ivana Malá.
20. ČABLA, A. (2012a). Předpovědi extrémních říčních průtoků metodou „peaks over threshold" a jejich kvalita*. Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze*. Oeconomica. Praha
21. ČABLA, A. (2012b). Unemployment duration in the Czech Republic. Prague 13.09.2012 – 15.09.2012. In: *The 6th International Days of Statistics and Economics, Conference Proceedings*. Praha, 2012, pp. 257 – 267. ISBN 978-80-86175-86-7.
22. ČABLA, A. (2015) Unemployment Duration in the Czech Republic After the Economic Crisis. In: *Applications of Mathematics and Statistics in Economics – AMSE* [CD ROM]. Jindřichův Hradec, 02.09.2015 – 06.09.2015. Praha : Oeconomica Publishing House, 2015. 11 s. ISBN 978-80-245-2099-5.
23. ČABLA, A. (2016) Minimal Adequate Model of Unemployment Duration in the Post-Crisis Czech Republic. *Statistika*. roč. 96, č. 1, s. 50–62. ISSN 0322-788X.
24. ČSÚ. MPSV. (2012). *Společná tisková zpráva Českého statistického úřadu a Ministerstva práce a sociálních věcí ČR: Změna výpočtu ukazatele registrované nezaměstnanosti* [online]. [cit. 2016-10-10].
25. ČSÚ. (2013). *Pokyny pro tazatele a krajské garanty pro rok 2014: Výběrové šetření pracovních sil.* Praha.
26. ČSÚ (2015) Výběrové šetření pracovních sil. *Český statistický úřad* [online]. [cit. 2016-10-10]. Dostupné z: https://www.czso.cz/csu/vykazy/vyberove_setreni_pracovnich_sil
27. ČSÚ (2016), Zaměstnanost a nezaměstnanost podle výsledků VŠPS - Metodika. *Český statistický úřad* [online]. [cit. 2016-10-10]. Dostupné z: https://www.czso.cz/csu/czso/zam_vsps

28. DAVISON, A. SMITH, R. (1990). Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society. Series B (Methodological), 52*(3), 393-442.
29. DE GRUTTOLA, V. LAGAKOS, S. (1989). Analysis of Doubly-censored Survival Data, with Application to AIDS. *Biometrics*. 45
30. DE HAAN, L. BALKEMA, A.A. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2.
31. DELIGNETTE-MULLER, M.R. DUTANG, C. (2015). fitdistriplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, 64(4).
32. DEMPSTER, A.P., LAIRD, N.M, RUBIN, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society, Series B*. 39.
33. DERKSEN, S. KESELMAN, H.J. (1992). Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables. *British Journal of Mathematical and Statistical Psychology*. 45.
34. DEY, D., YAN, J. (2016) *Extreme Value Modeling and Risk Analysis: Methods and Applications*. Boca Raton, FL: CRC Press, Taylor. ISBN 1498701299.
35. EFRON, B. (1967). The Two Sample Problem with Censored Data. In*: Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability*. New York: Prentice-Hall.
36. ESSER, M., POPELKA, J..(2003) Analysis of Factors Influencing Time of Unemploment Using Survival Time Analysis. Bratislava 04.12.2003 – 05.12.2003. In: *Výpočtová štatistika*. Bratislava: Slovenská štatistická a demografická spoločnosť, 2003, s. 250–254. ISBN 80-88946-29-8.
37. EUROSTAT. (2016). European Union Labour Force Survey (EU LFS). *Eurostat* [online]. 2016 [cit. 2016-10-10]. Dostupné z: http://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey
38. MICHAEL, F. (2013) Reinventing Pareto: Fits for both small and large losses. In: *ASTIN Colloquium* [online]. The Hague, s. 24 [cit. 2016-12-09]. Dostupné z: http://www.actuaries.org/ASTIN/Colloquia/Hague/Papers/Fackler.pdf
39. FAY M.P. (1996). Rank Invariant Tests for Interval Censored Data under the Grouped Continuous Model. *Biometrics*, **52**,
40. FAY M.P. (1999). Comparing Several Score Tests for Interval Censored Data. *Statistics in Medicine*, **18**.

41. FAY, M.P., SHAW, P.A. (2011). Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package. *Journal of Statistical Software*, **36**(2), 1-34. URL http://www.jstatsoft.org/v36/i02/.
42. FERREIRA, A. DE HAAN, L. (2015) On the block maxima method in extreme Odhad theory: PWM estimators. *The Annals of Statistics*. 43(1), 276-298. DOI: 10.1214/14-AOS1280. ISSN 0090-5364.
43. FINKELSTEIN, D.M. (1986). A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics*, 42.
44. FISHER, R. A., L. H. C. TIPPETT. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180-190, DOI: 10.1017/S0305004100015681.
45. FRÉCHET M.R. (1927). Sur la loi de probabilité de lécart maximum. *Polonaise de Mathématique*, 6, 93-116.
46. FRIGESSI, A., HAUG, O. HAVARD, R. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection, *Extremes*, 5, 219–235.
47. GELMAN, A., SU, Y. (2016). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. R package version 1.9-3. https://CRAN.R-project.org/package=arm
48. GENTLEMAN, R. a C.J. GEYER. (1994). Maximum Likelihood for Interval Censored Data: Consistency and Computation. *Biometrica*. **81**(3).
49. GENTLEMAN, R. a A.C. VANDAL. (2001).Computational Algorithms for Censored-Data Problems Using Intersection Graphs. *Journal of Computational and Graphical Statistics*. **10**(3).
50. GILL, R.D. (1980). Censoring and Stochastic Integrals. *Mathematical Centre Tracts*. Amsterodam: Mathematisch Centrum. 124.
51. GNĚDĚNKO, B.V. (1943). Sur la distribution limite du terme maximum dune série aléatorie. *Annals of Mathematics*, 44, 423-53
52. GOGGINS, W., FINKELSTEIN, D., SCHOENFLED, D. ZASLAVSKY, A. (1998). A markov Chain Monte Carlo EM Algorithm for Analyzing Interval Censored Data under the Cox Proportional Hazards Model. *Biometrics*. 54.
53. GRAUNT, J. *Natural and Political Observations Made upon the Bills of Mortality*. Baltimore: Johns Hopkins Press, 1939. (původní vydání 1662).
54. GRAMBSCH, P.M., O´Brien, P.C. (1991). The Effects of Transformations and Preliminary Tests for Non-linearity in Regression. *Statistics in Medicine*. 10.

55. GROENEBOOM, P., WELLNER, J. (1992). *Information Bounds and Non-parametric Maximum Likelihood Estimation*. DMV Seminar, Band 19, Birkhauser, New York.
56. HARRELL JR., F. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer, 2001. ISBN 978-0387-95232-1.
57. HOSKING, J. WALLIS, J. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics, 29*(3), 339-349.
58. HOSMER, D.W., LEMESHOW, S.. MAY, (1999). S. *Applied survival analysis: regression modeling of time to event data*. New York: John Wiley & Sons. Wiley series in probability and statistics. ISBN 0-471-15410-5.
59. ILO. (1982) *Thirteenth International Conference of Labour Statisticians: Resolution concerning statistics of the economically active population, employment, unemployment and underemployment*.
60. JAROŠOVÁ E., MALÁ I., ESSER M., POPELKA J. (2004). COMPSTAT 2004 Symposium: 1 – 8.
61. JAROŠOVÁ, E., MALÁ I. (2005) "Modelling time of unemployment in the Czech Republic." In: KOVÁČOVÁ, M. *Proceedings of the 4th International Conference APLIMAT 2005*. 2005. Bratislava: Slovak University of Technology. ISBN 978-80-969264-2-8.
62. JARUŠKOVÁ, Daniela; HANEK, Martin (2006) Peaks over threshold method in comparison with block-maxima method for estimating high return levels of several Northern Moravia precipitation and discharges series. *Journal of Hydrology and Hydromechanics*, 54 (4). pp. 309-319. ISSN 0042-790X
63. JONGBLOED, G. (1998). The Iterative Convex Minorant Algorithm fot Nonparametric Estimation, *Journal of Computational and Graphical Statistics*, 7.
64. KALBFLEISCH, J. LAWLESS, J. (1985). The Statistical Analysis of Panel Data under a Markov Assumption. *Journal of American Statistical Association*. 80.
65. KAPLAN, E. MEIER, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. 53
66. KERCKHOFFS, C., DE NEUBORUG, C., PALM, F. (1994). The determinants of unemployment and job search duration in the Netherlands. *De Economist*, 142(1), 21-42.

67. KIEFER, J., WOLFOWITZ, J. (1956) Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters, *Ann. Math. Statist.* 27.
68. KIESELBACH, T. (2003) Long-term unemployment among young people: the risk of social exclusion. *American journal of community psychology*. ISSN 0091-0562.
69. KLEIN, J.P. (1991). Small-Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators. *Scandinavian Journal of Statistics*. 18.
70. KLEIN, J.P., MOESCHBERGER, M.L. (1984). The Asymptotic Bias of the Product-Limit Estimator Under Dependent Competing Risks. *Indian Journal of Productivity, Reliability and Quality Control*.
71. KLEIN, J. P., MOESCHBERGER, M. L. (1997). *Survival Analysis : Techniques for Censored and Truncated Data*. New York : Springer-Verlag. 502 s. ISBN 0-387-94829-5.
72. KLEINBAUM, D. G., KLEIN, M. (2012). *Survival analysis: a self-learning text*. 3rd ed. New York: Springer. Statistics for biology and health. ISBN 14-419-6646-3.
73. KOTÝNKOVÁ, M. (2006). Long-Term Unemployment in the Czech Republic: Motivation, Obstacles and the Social Assistance System. *Prague Economic Papers*. **15**(2), 99-112. DOI: 10.18267/j.pep.279. ISSN 1210-0455.
74. KUHLEKASPER, T. STEINHARDT, M. (2011). Unemployment Duration in Germany – A comprehensive study with dynamic hazard models and P-Splines. *Hwwi Research*, 111. Hamburg Institute of International Economics.
75. LANDER, J.P. (2016), *coefplot: Plots Coefficients from Fitted Models. R package version 1.2.4.* https://CRAN.R-project.org/package=coefplot
76. LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Second Edition. Hoboken (NJ) : John Wiley & Sons. 630 s. ISBN 0-471- 37215-3.
77. LAW, C., BROOKMEYER, R. (1992). Effects of Mid-Point Imputation on the Analysis of Doubly Censored Data. *Statistics in Medicine*, 11.
78. LAYARD, P. R. G. NICKELL, S. J. JACKMAN, R. (1991) *Unemployment: macroeconomic performance and the labour market*. New York: Oxford University Press. ISBN 01-982-8434-9.
79. LIN, Ch. WANG, W. (2007). Estimation for the generalized pareto distribution with censored data. *Communications in Statistics - Simulation and Computation* DOI: 10.1080/03610910008813660.

80. LIU, X. (2012). *Survival analysis: models and applications*. Peking: Higher Education Press. ISBN 04-709-7715-9.
81. MACDONALD, A., SCARROTT, C.J., LEE D., DARLOW B., REALE M., RUSSELL G. (2011) A Flexible Extreme Value Mixture Model. *Computational Statistics & Data Analysis*. **56**(6), 21.
82. MALÁ, I. (2014) The Use of Finite Mixture Model for Describing Differences in Unemployment Duration. In: *AMSE* [CD ROM]. Jerzmanovice, 27.08.2014 – 31.08.2014. Wroclaw : Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu. s. 164–172. ISBN 978-83-7695-421-9.
83. MALÁ, I. (2015) Modelling of unemployment duration in the Czech Republic with the use of L-moments. In: *Applications of Mathematics and Statistics in Economics – AMSE* [CD ROM]. Jindřichův Hradec, 02.09.2015 – 06.09.2015. Praha : University of Economics, Prague, Oeconomica Publishing House. 8 s. ISBN 978-80-245-2099-5.
84. MANTEL, N. (1970). Why Stepdown Procedures in Variable Selecetion. *Technometrics*. 12.
85. MICKEY, J., GREENLAND S. (1989). A Study of the Impact of Confounder Selection Criteria on Effect Estimation. *American Journal of Epidemiology*. 129.
86. MOSSAKOWSKI, K. (2009) The Influence of Past Unemployment Duration on Symptoms of Depression Among Young Women and Men in the United States: a longitudinal analysis covering 63 countries, 2000–11. *American Journal of Public Health*. **99**(10), 1826-1832. DOI: 10.2105/AJPH.2008.152561. ISSN 0090-0036.
87. MUSSIDA, C. (2007). Unemployment duration and competing risks: A regional investigation. *Quaderni del Dipartimento di Scienze Economiche e Sociali*, 0749. Università Cattolica del Sacro Cuore, Dipartimenti e Istituti di Scienze Economiche.
88. NELSON, W. (1972). Theory and Applications of Hazard Plotting for Censored Failure Data. *Technometrics*. 14.
89. NORDT, C. WARNKE, I. SEIFRITZ, E. KAWOHL, W. (2015) Modelling suicide and unemployment: a longitudinal analysis covering 63 countries, 2000–11. *The Lancet Psychiatry*. **2**(3), 239-245. DOI: 10.1016/S2215-0366(14)00118-7. ISSN 22150366.
90. OECD. (2016a). Average Duration of Unemployment. *OECD.Stat* [online]. [cit. 2016-10-10]. Dostupné z: https://stats.oecd.org/Index.aspx?DataSetCode=AVD_DUR
91. OECD. (2016b). *LABOUR FORCE STATISTICS IN OECD COUNTRIES: SOURCES, COVERAGE AND DEFINITIONS* [online]. 39 s. [cit. 2016-10-10]. Dostupné z: http://www.oecd.org/els/emp/LFSNOTES_SOURCES.pdf.

92. PARODI, P. (2014). *Pricing in general insurance*. London: CRC Press, Taylor. ISBN 978-146-6581-449.
93. PEDACE, R. ROHN, S. (2011) The Impact of Minimum Wages on Unemployment Duration: Estimating the Effects Using the Displaced Worker Survey. *Industrial Relations: A Journal of Economy and Society*. **50**(1), 57-75. DOI: 10.1111/j.1468-232X.2010.00625.x. ISSN 00198676.
94. PETERS, G. W., SHEVCHENKO, P.V. (2015) *Advances in heavy tailed risk modeling: a handbook of operational risk*. Hoboken, New Jersey: Wiley. Wiley handbooks in financial engineering and econometrics. ISBN 978-1-118-90953-9.
95. PETO, R., PETO, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society A*, **135**.
96. PETO, R. (1973). Experimental Survival Curves for Interval-Censored Data. *Applied Statistics*. 22.
97. PICKANDS, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3.
98. PISSARIDES, C. A. (1992): Loss of Skill during Unemployment and the Persistence of Employment Shocks. *Quarterly Journal of Economics*, 107, č. 4, str. 1371–1391
99. PORTUGAL, P. (2007). U.S. Unemployment Duration: Has Long Become Bonger or Short Become Shorter? *Working Paper Series*. Praha: Czech National Bank, (17). ISSN 1803-2397.
100. R CORE TEAM. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
101. ROTTE, R., STEININGER. M. (2008). Crime, Unemployment, and Xenophobia? An Ecological Analysis of Right-Wing Election Results in Hamburg, 1986−2005. *IZA Discussion Papers*. ISSN 2365-9793.
102. SATTEN, G. (1996). Rank-based Inference in the Proportional Hazards Model for Interval-censored Data. *Biometrika*. 83.
103. SCARROTT, C., MACDONALD A. (2012). A Review of Extreme Value Threshold Estimation and Uncertainity Quantification. *REVSTAT – Statistical Journal*. **10**(1), 28.
104. SCHWARZ, G.E. (1978). Estimating the Dimension of a Model. *Annals of Statistics*. 6(2). 461 – 464.

105. SMITH, R. (1985). Maximum Likelihood Estimation in a Class of Nonregular Cases. *Biometrika*, *72*(1), 67-90.

106. SO, Y., JOHNSTON, G., KIM, S.H. (2010), "Analyzing Interval-Censored Survival Data with SAS® Software," *Proceedings of the SAS® Global Forum 2010 Conference*, Cary, NC: SAS Institute Inc.

107. SUN, J. (1996). A non-parametric test for interval-censored failure time data with application to AIDS studies.*Statistics in Medicine*. 15(13).

108. SUN, J. (2006) *The statistical analysis of interval-censored failure time data*. New York: Springer. ISBN 03-873-2905-6.

109. TANAKA, S. TAKARA, A. (2002). A study on threshold selection in POT analysis of extreme floods. *The Extremes of the Extremes: Extraordinary Floods.* Oxfordshire, UK: IAHS Publ. 299 - 304.

110. THEODOSSIOU, I. YANNOPOULOS, A. (1998). Labour market segmentation and unemployment duration. *Applied Economics Letters*, 5 (9).

111. THEODOSSIOU, I. ZAROTIADIS, G. (2010). Employment and unemployment duration in less developed regions. *Journal of Economic Studies*, 37(5), 505 – 524.

112. THERNEAU, T. (2015) *A Package for Survival Analysis in S*. version 2.38. http://CRAN.R-project.org/package=survival.

113. THERNEAU, T. GRAMBSCH, P. (2000). *Modelling Survival Data: Extending the Cox Model.* Springer, New York, ISBN 0-387-98784-3.

114. TURNBULL, B.W. (1974). Nonparametric Estimation of Survivorship Function with Doubly Censored Data. *Journal of the American Statistical Association*. 69.

115. TURNBULL, B.W. (1976). The Empirical Distribution Function with Arbitrary Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society: Series B.*, 69.

116. VOJTĚCH, J. (2010). Možnosti využití teorie extrémních hodnot. Sborník prací účastníků vědeckého semináře doktorandského studia Fakulty informatiky a statistiky VŠE v Praze. Oeconomica. Praha.

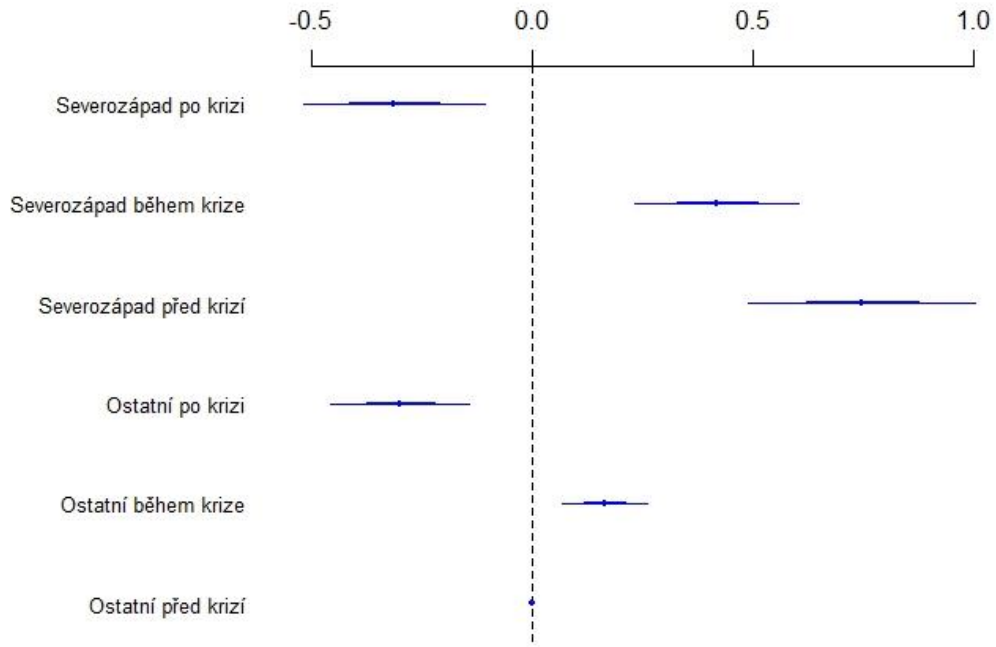
117. VOJTĚCH, J. (2011). *Využití teorie extrémních hodnot při řízení operačních rizik*. Praha. Disertační práce. Vysoká škola ekonomická v Praze. Vedoucí práce Jana Kahounová.

118. WALD, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*. 54.

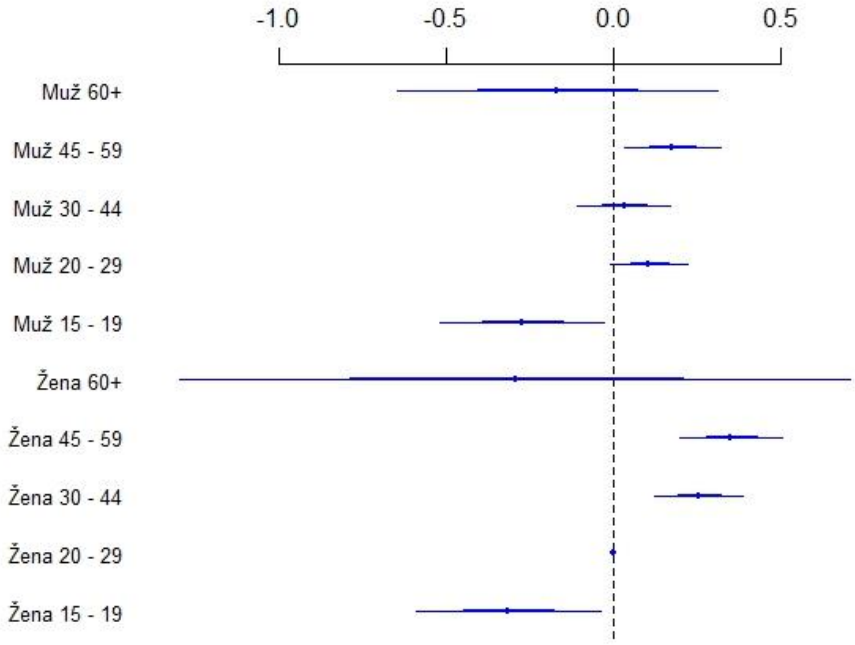
119. WELLNER, J., ZHANG Y. (1997). A Hybrid Algorithm fot Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data. *Journal of the American Statistical Association*. 92.

120. WUERTZ, D. a další, *fExtremes: Rmetrics – Extreme Financial Market Data*. R package verze 3010.81. Dostupné z: https://CRAN.R-project.org/package=fExtremes

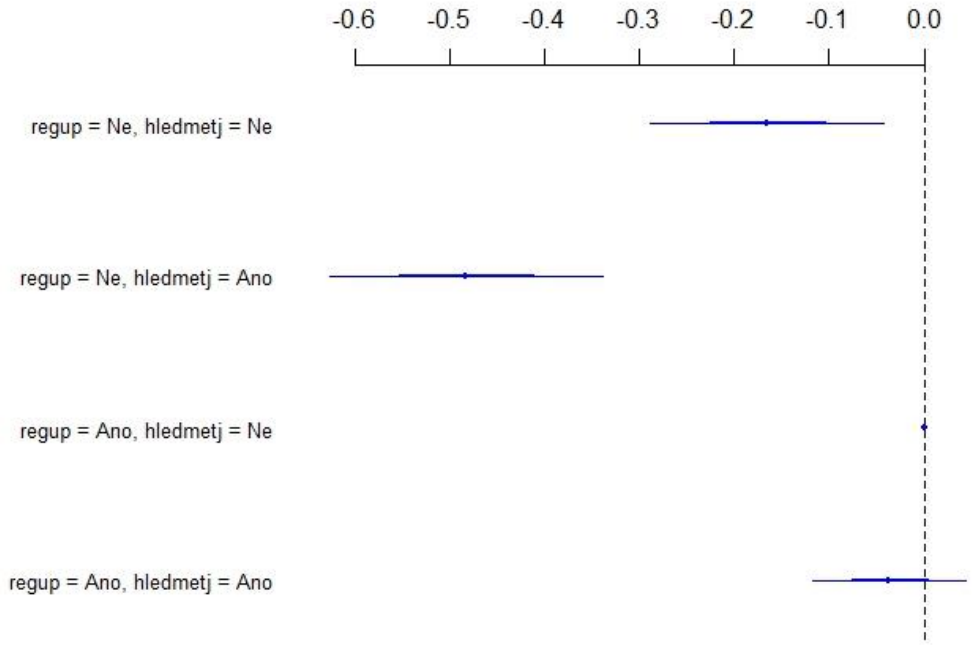
Příloha



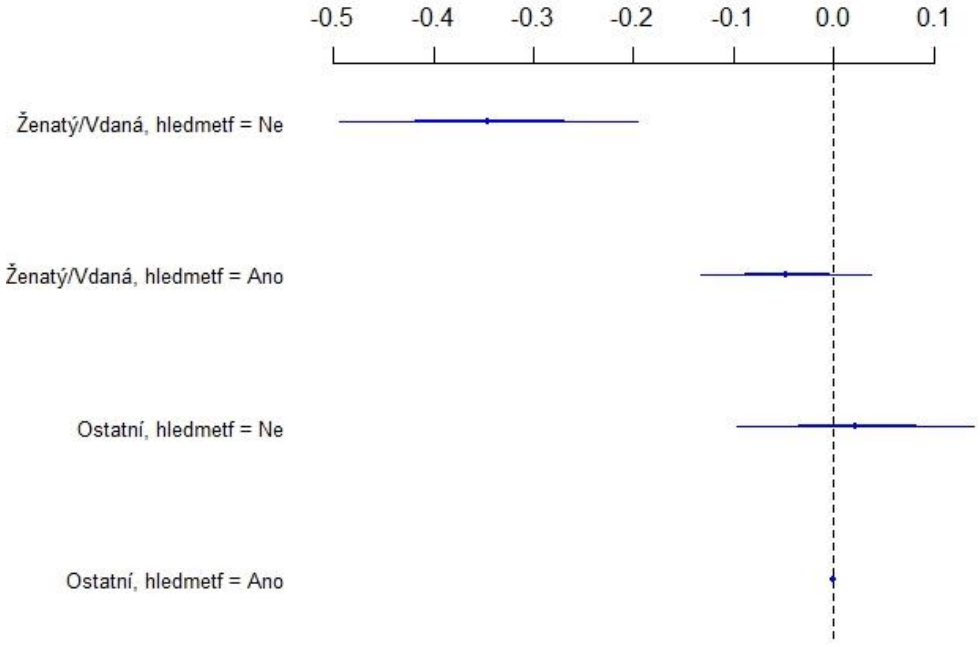
Obrázek 1



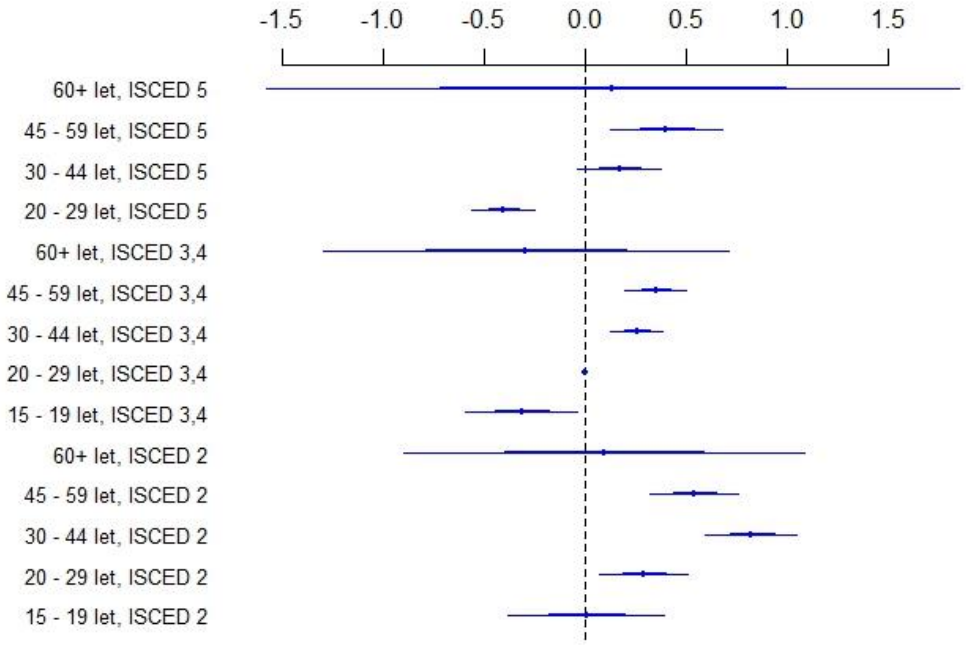
Obrázek 2



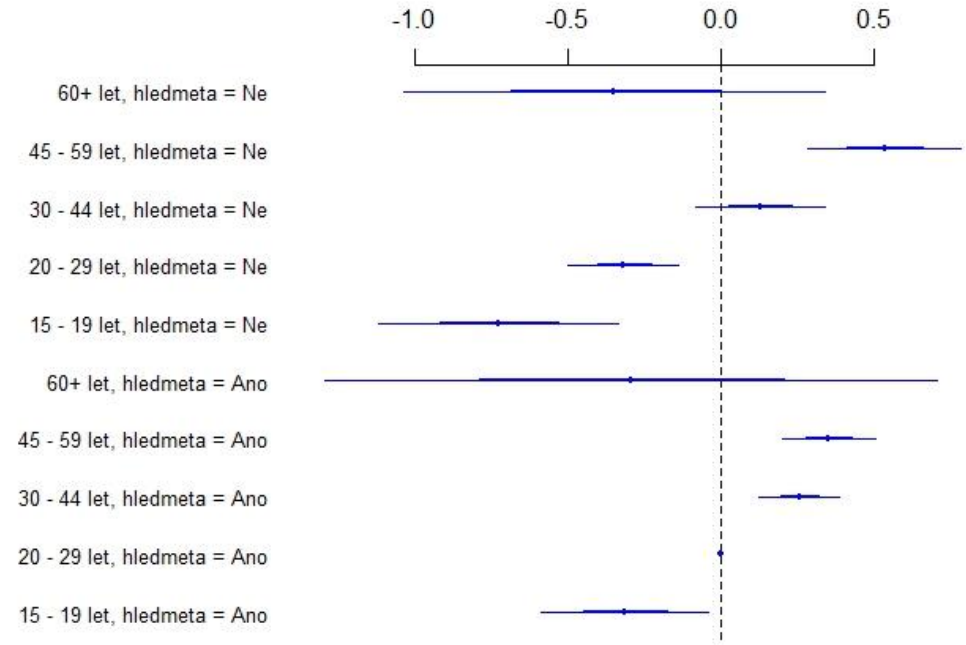
Obrázek 3



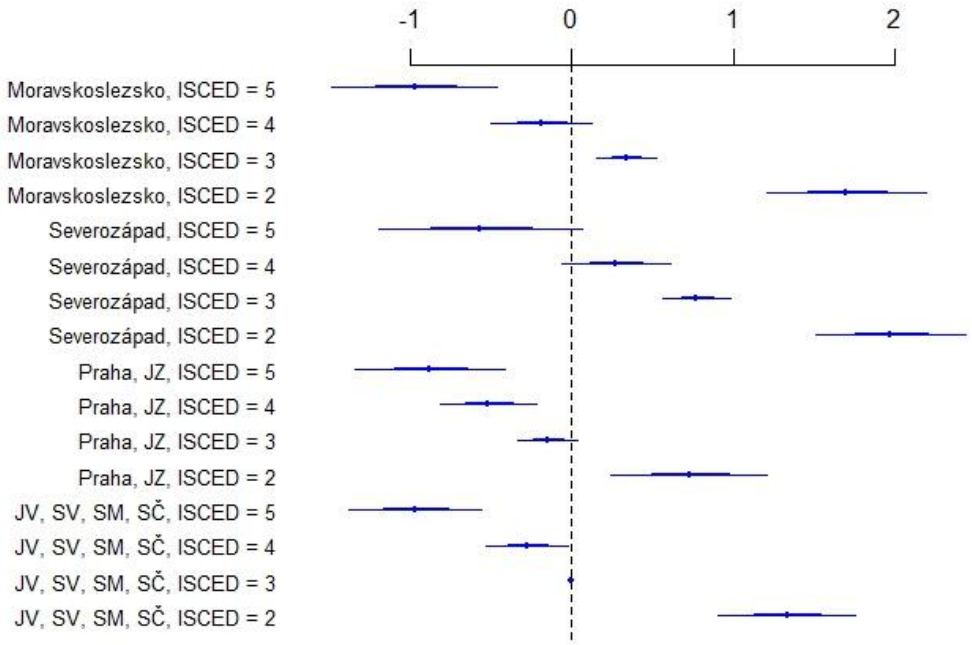
Obrázek 4



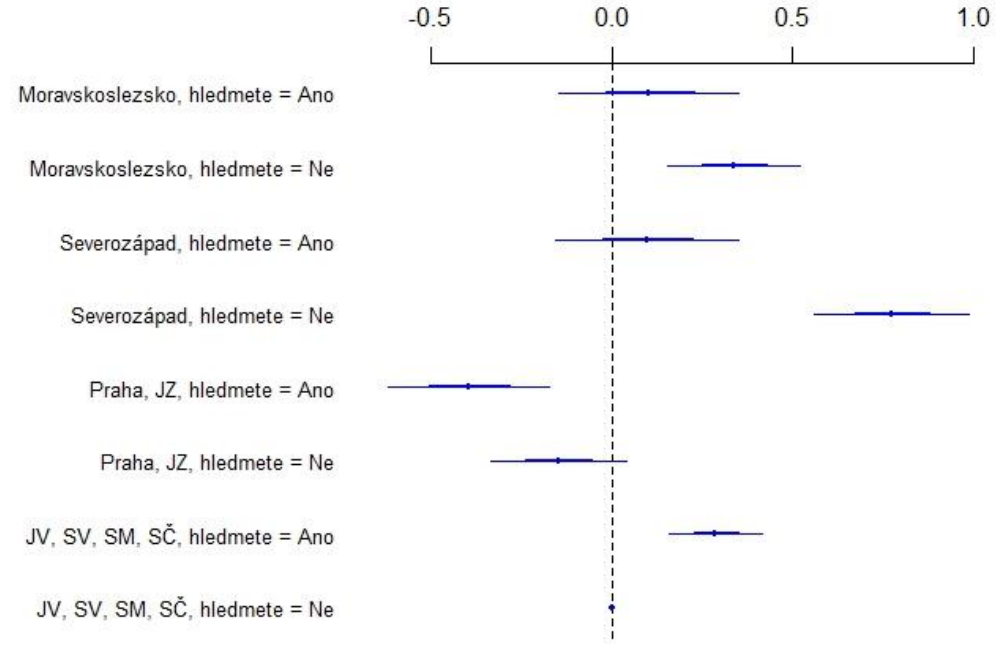
Obrázek 5



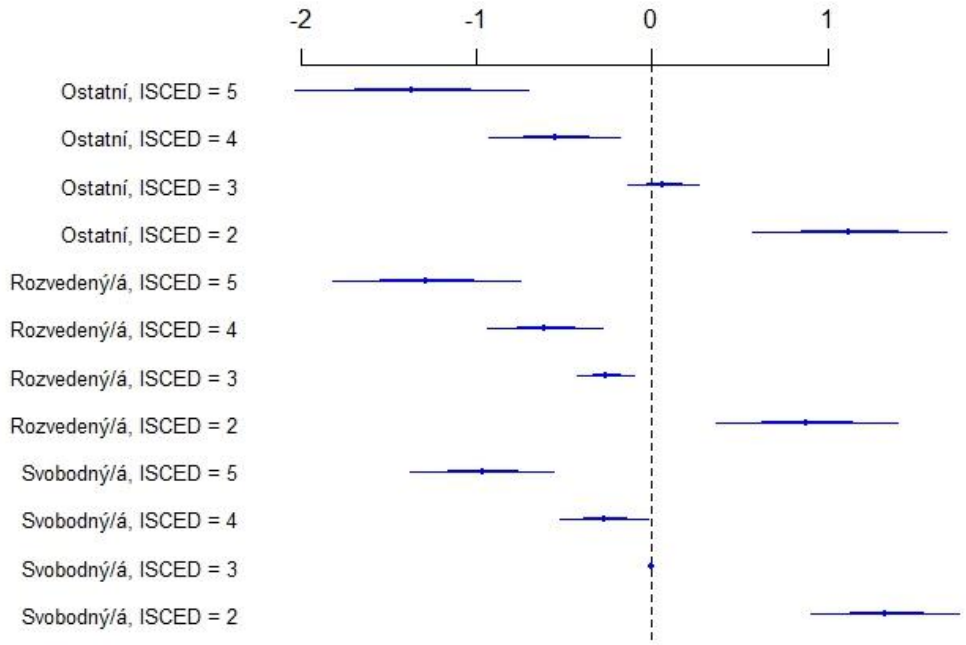
Obrázek 6



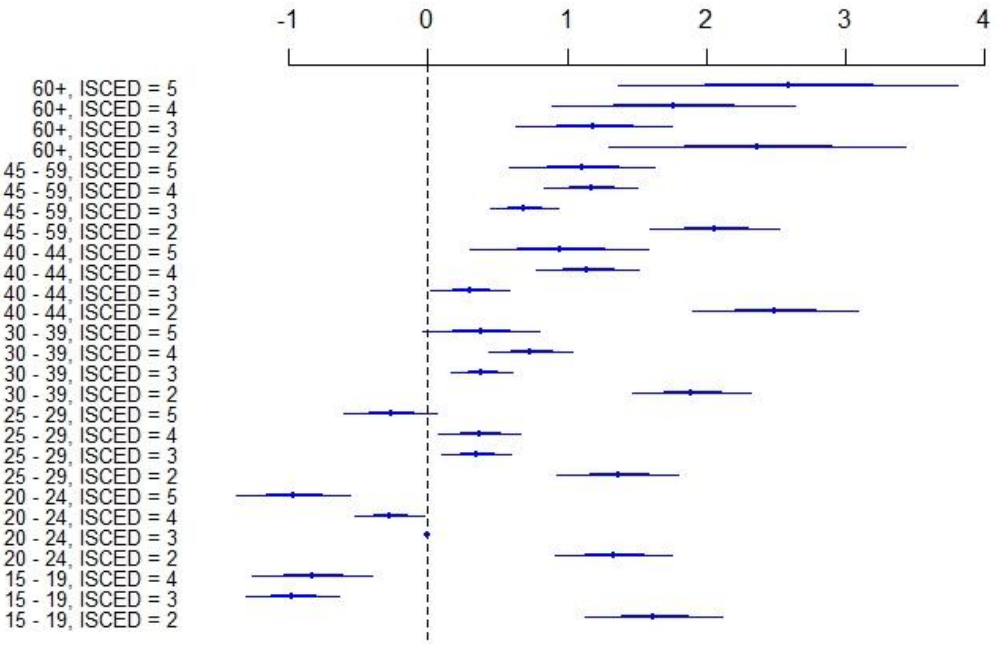
Obrázek 7



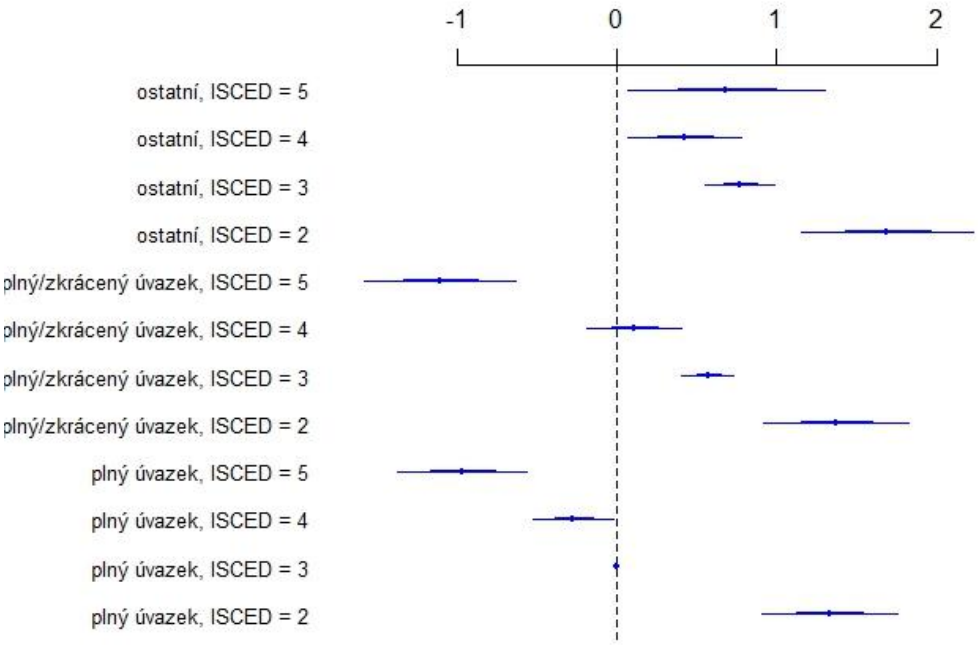
Obrázek 8



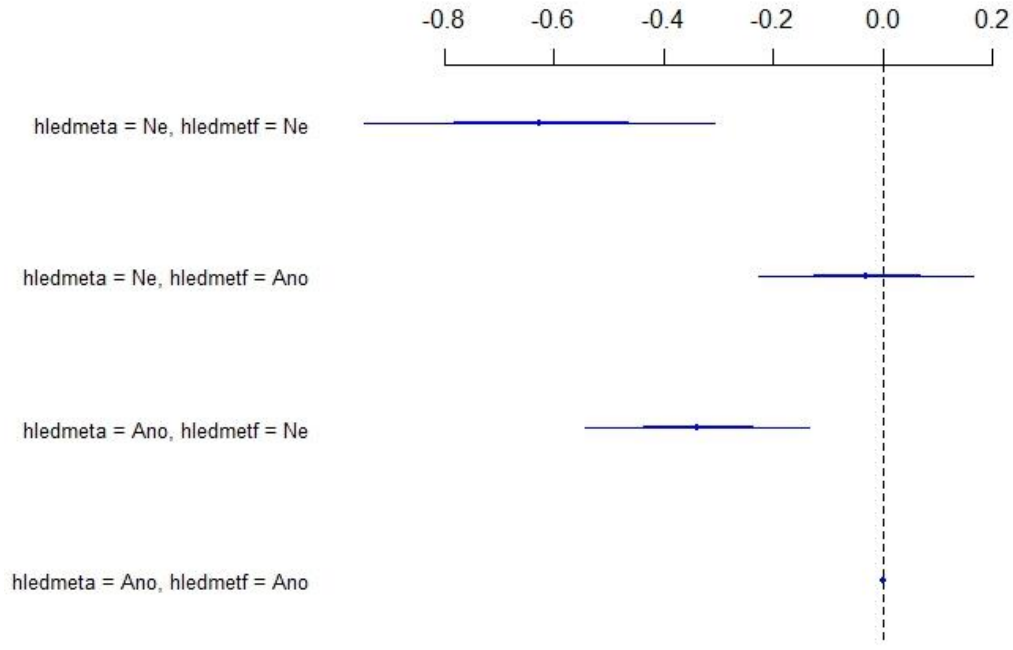
Obrázek 9



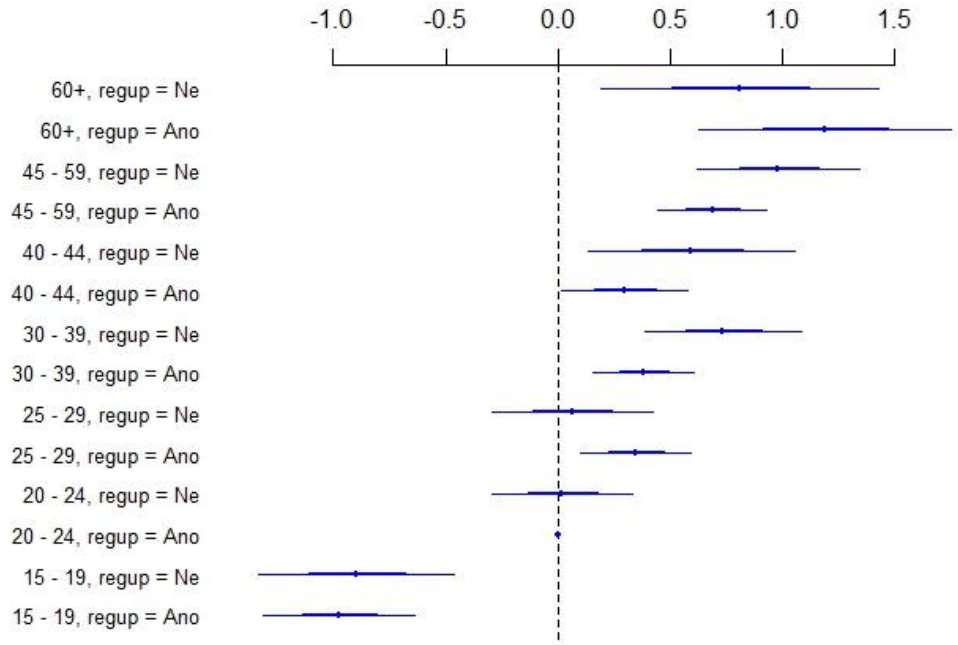
Obrázek 10



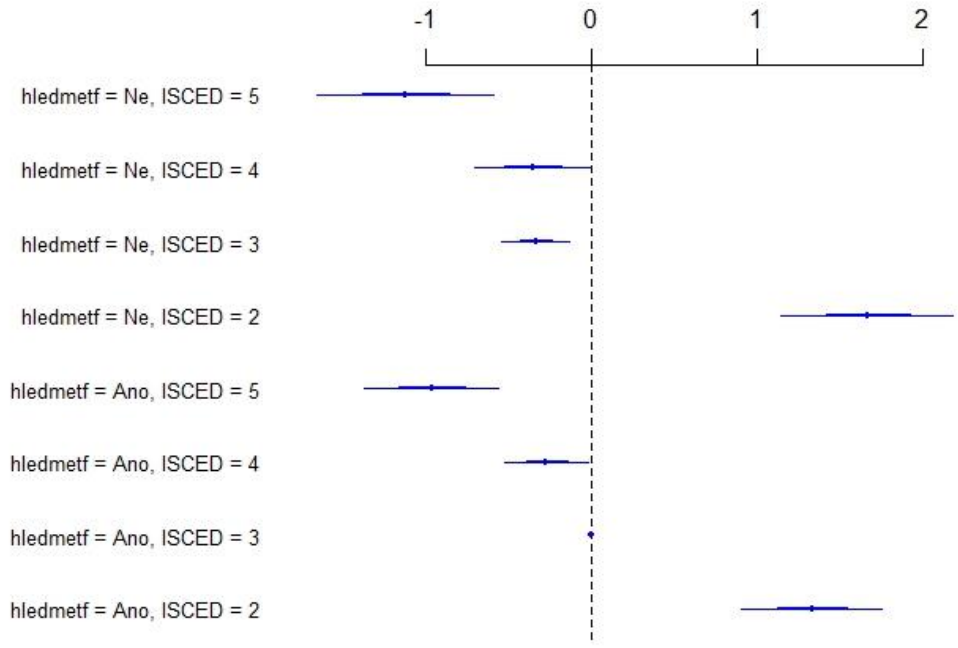
Obrázek 11



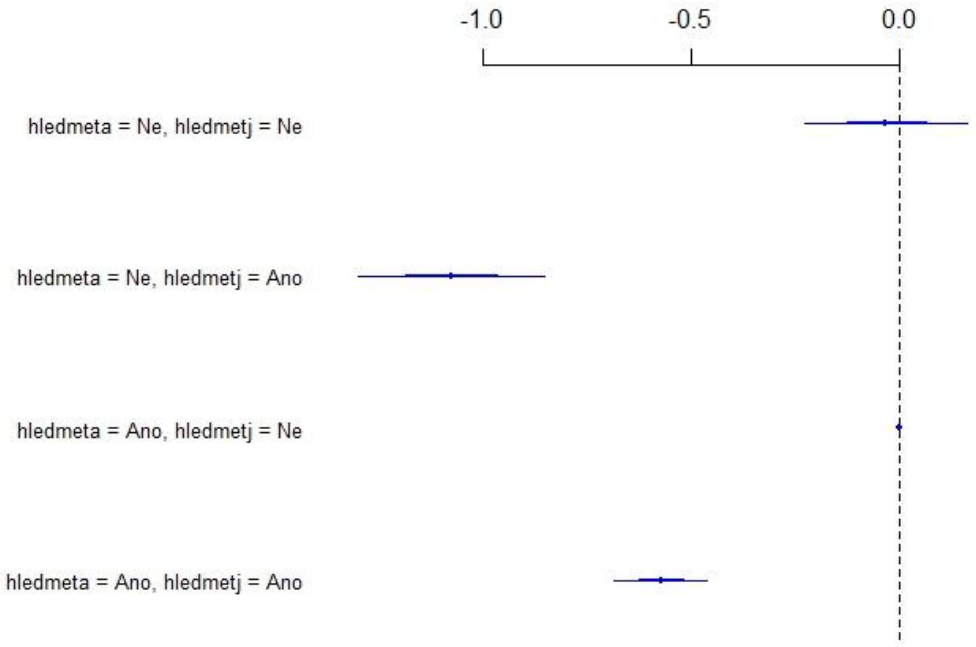
Obrázek 12



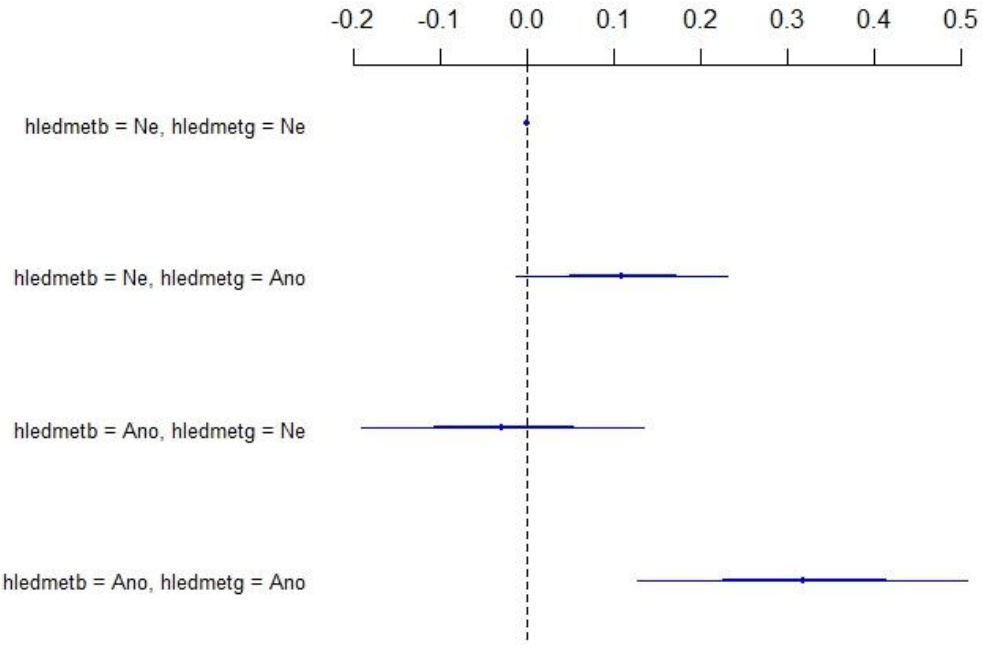
Obrázek 13



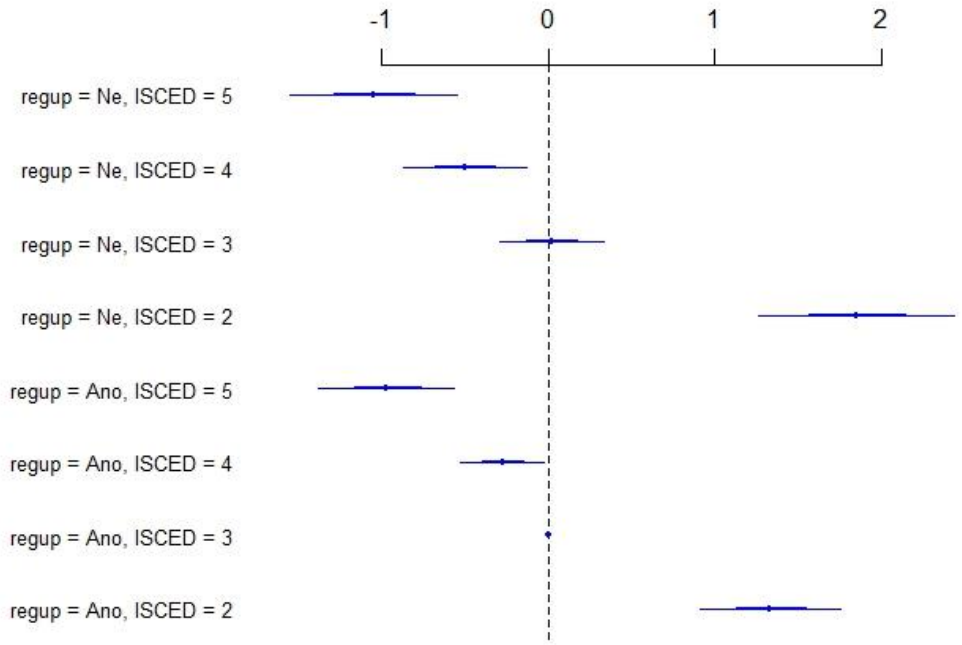
Obrázek 14



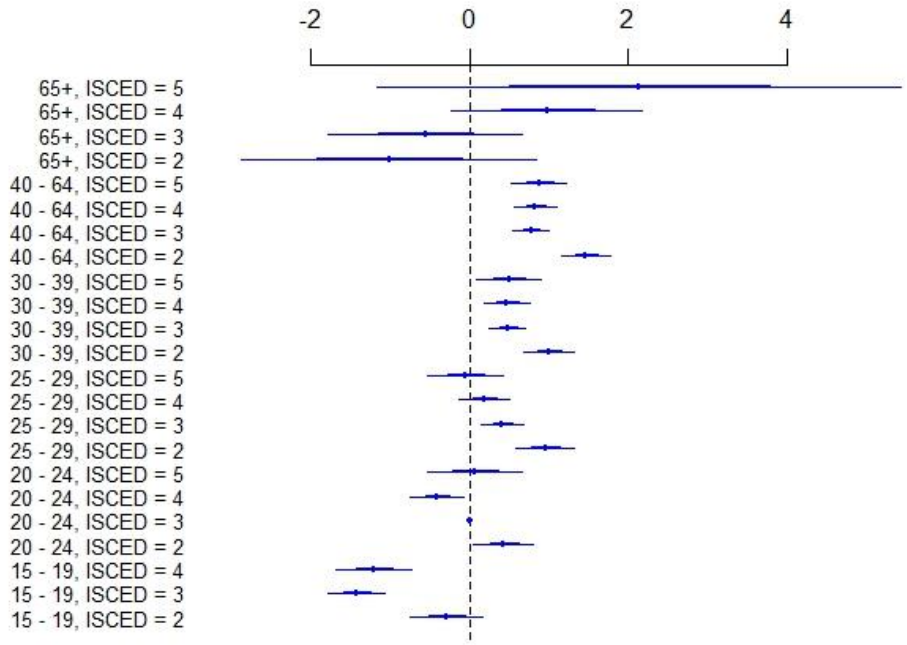
Obrázek 15



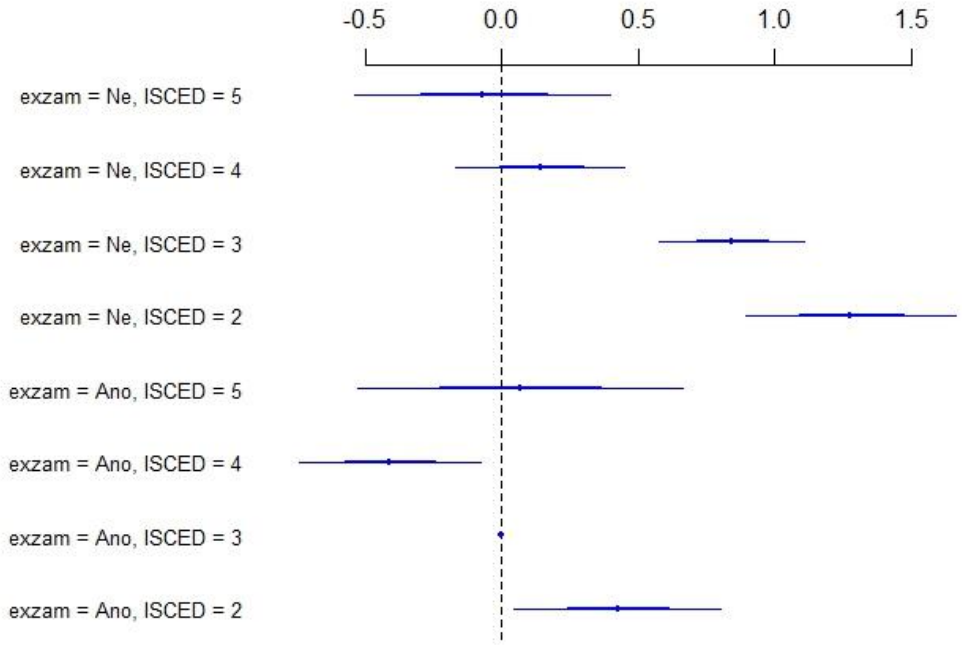
Obrázek 16



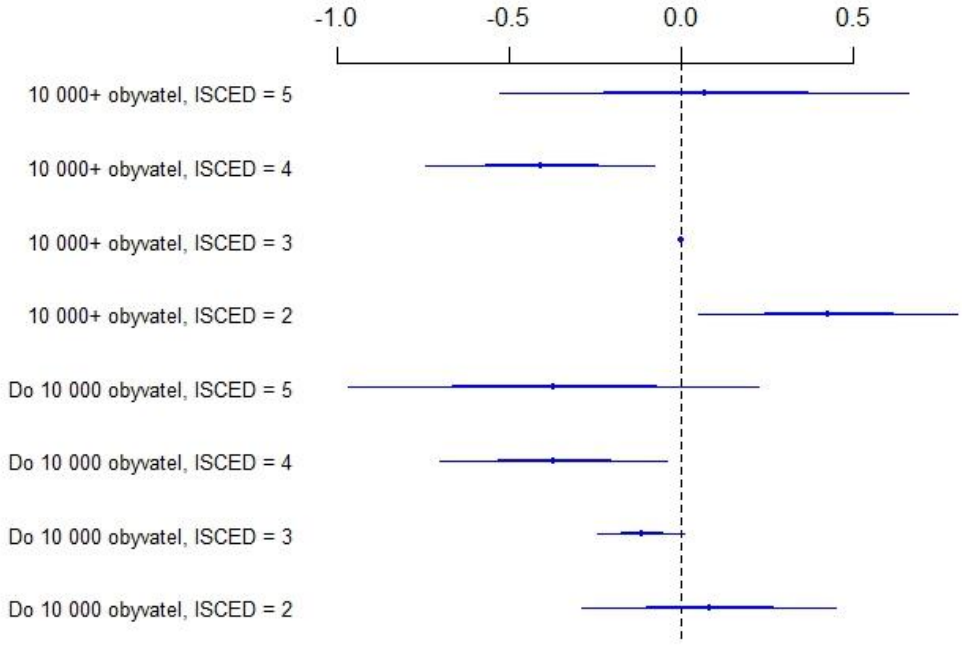
Obrázek 17



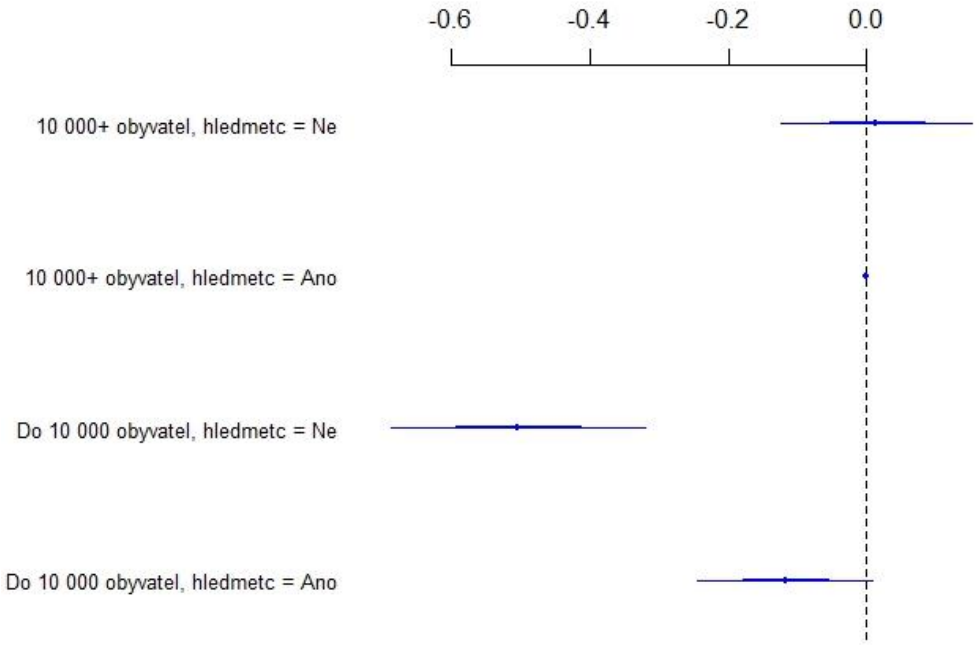
Obrázek 18



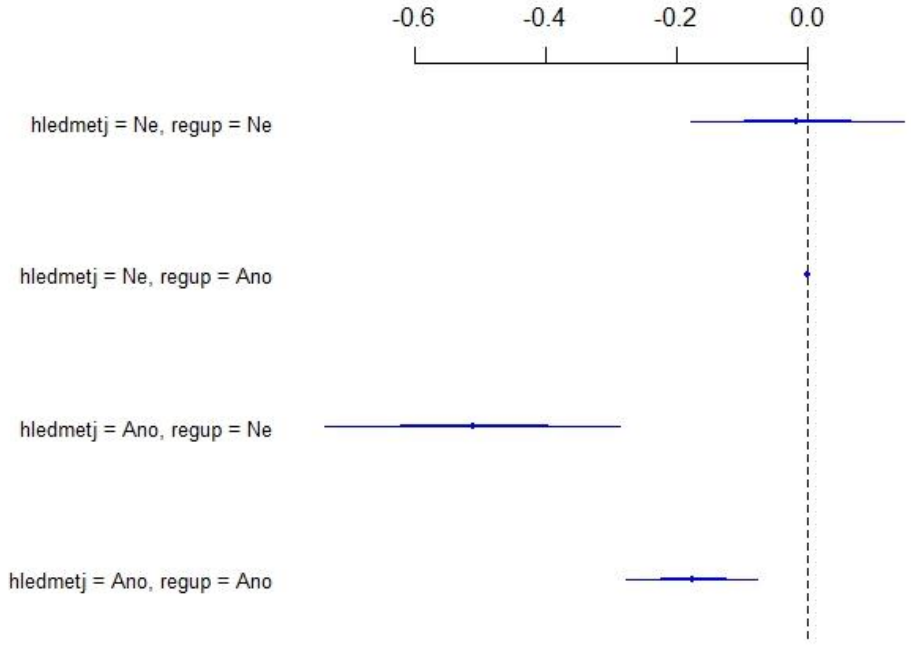
Obrázek 19



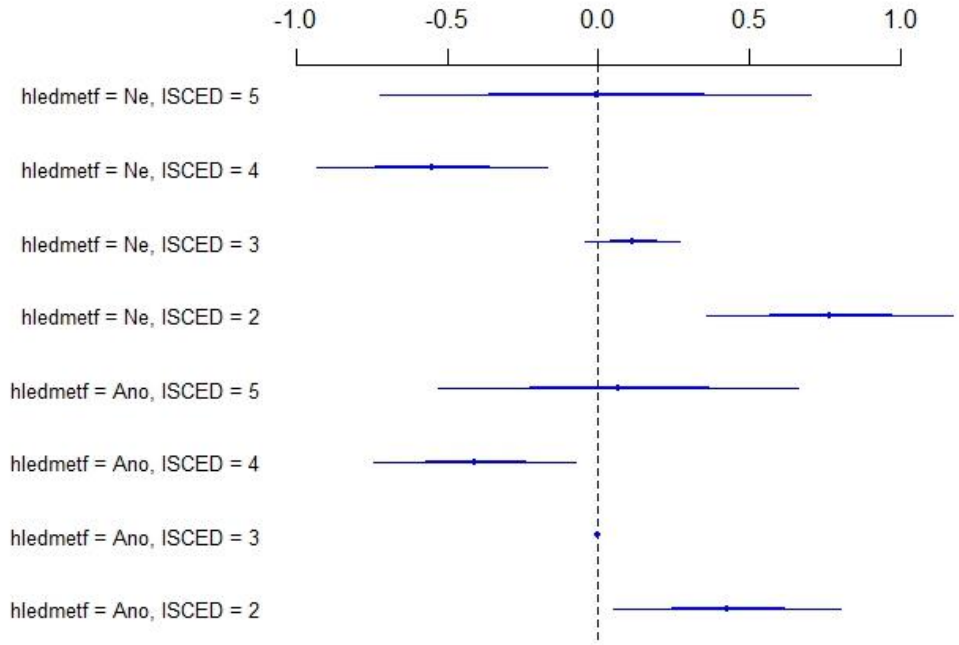
Obrázek 20



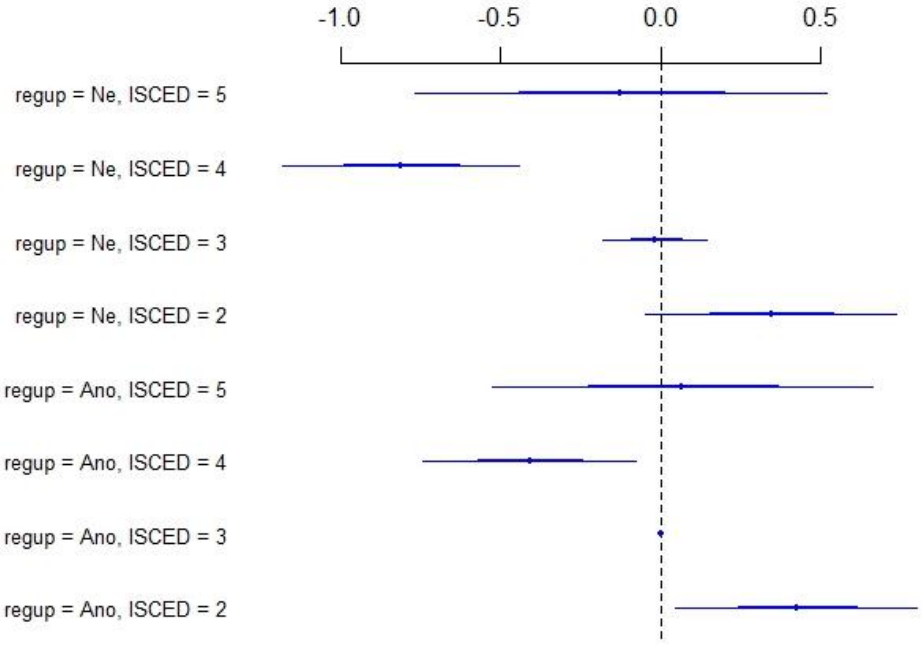
Obrázek 21



Obrázek 22



Obrázek 23



Obrázek 24

Seznam použitých znaků v pořadí výskytu

N **–** počet nezaměstnaných

Zam – počet zaměstnaných

EA – počet ekonomicky aktivních osob

OBN – obecná míra nezaměstnanosti

L – spodní hranice intervalu

R – horní hranice intervalu

X, T – zkoumaná náhodná veličina

x, t – hodnota zkoumané náhodné veličiny

$S(t)$ – funkce přežití

$F(t)$ – distribuční funkce

$h(t)$ – riziková funkce

$f(t)$ – hustota pravděpodobnosti

$H(t)$ – kumulovaná riziková funkce

C – čas pozorování

I – indikátorová funkce

δ – informace o cenzorování

W, V – čas pozorování

X, S – čas výskytu

i – pořadí jednotky

$\boldsymbol{\theta}$ – obecně značený vektor parametrů

$L()$ – věrohodnostní funkce

$l()$ – logaritmus věrohodnostní funkce

n – počet pozorování

j – pořadí časového bodu

m – počet časových bodů

τ – časový bod

α_{ij} – váha pro výpočet Turnbullova odhad

d – počet událostí

Y – počet jednotek v riziku

u_j – Lagrangeovy multiplikátory

x – vektor vysvětlujících proměnných

β – vektor parametrů příslušné k vektoru proměnných **x**

r **–** pořadí proměnné

γ – obecně vektor parametrů funkce přežití

ε – chybová složka regresního modelu

$g()$ – monotónní transformace

U –skóry

l – pořadí řádku ve vektoru skórů **U**

V – kovarianční matice

w – vektor vah

Q – testové kritérium log-rank testu

k – počet funkcí přežití

k – vektor konstant

Z – testové kritérium s (asymptoticky) normovaným normálním rozdělením

μ – parametr polohy

σ – parametr měřítka

I – informační matice

Λ – testové kritérium testu věrohodnostním poměrem

p – počet parametrů

M – počet stupňů volnosti v testu věrohodnostním poměrem

α, β, β^* – parametry log-logistického rozdělení

q – počet náhodných veličin

$G()$ – distribuční funkce zobecněného rozdělení extrémní hodnoty

$\Phi()$, $\Psi()$, $\Lambda()$ – distribuční funkce Fréchetova, Weibullova a Gumbellova rozdělení

ξ – parametr konce rozdělení (zobecněného rozdělení extrémní hodnoty a zobecněného Paretova rozdělení)

u – práh

ω_F – hranice zobecněného Paretova rozdělní

$Z()$ – distribuční funkce zobecněného Paretova rozdělení

μ^* - dolní nosič zobecněného Paretova rozdělení

φ_u – podíl pravého konce rozdělení

$v()$ – hustota pravděpodobnosti směsi rozdělení

$p()$ – dynamická váha

$z()$ – hustota pravděpodobnosti zobecněného Paretova rozdělení