MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

DIPLOMOVÁ PRÁCE

Zuzana Skřivanová

Stochastické modelování úmrtnosti pro více populací

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	RNDr. Lucie Mazurová, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční a pojistná matematika

Praha 2016

Prohlašuji, že jsem tuto diplomovou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 28. 7. 2016 Podpis autora

Název práce: Stochastické modelování úmrtnosti pro více populací

Autor: Zuzana Skřivanová

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: RNDr. Lucie Mazurová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá možnostmi modelování a předpovídání věkově specifické míry úmrtnosti. Úvodní část práce shrnuje základní pojmy z oblasti demografie související s úmrtností a vymezuje základní přístupy k jejímu modelování. Následně jsou podrobně popsány tři nejpoužívanější stochastické modely – Lee-Carterův, Renshaw-Habermanův a Cairns-Blake-Dowdův. Stěžejní část práce se zabývá možnostmi využití těchto modelů pro modelování úmrtnosti současně v několika korelovaných populacích. Uvedené teoretické podklady jsou v závěrečné části práce numericky ilustrovány na úmrtnostních modelech pro populace České a Slovenské republiky.

Klíčová slova: věkově specifická míra úmrtnosti, stochastické modely úmrtnosti, modelování úmrtnosti v korelovaných populacích

Title: Stochastic mortality modelling for multiple populations

Author: Zuzana Skřivanová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Lucie Mazurová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This thesis deals with the possibilities of modelling and forecasting of age-specific mortality rates. The introductory part summarizes the basic terms from demography, which are related to mortality, and specifies elementary approaches to the mortality modelling. Subsequently there are in detail described the three most commonly used stochastic mortality models – Lee-Carter, Renshaw--Haberman and Cairns-Blake-Dowd. The fundamental part of this thesis deals with the possibilities of using these models for mortality modelling simultaneously in correlated populations. These theoretical bases are in the final part of this thesis numerically illustrated on the mortality models for populations of Czech and Slovak Republic.

Keywords: age-specific mortality rate, stochastic mortality models, mortality models for correlated populations

Na tomto místě bych ráda poděkovala vedoucí své diplomové práce RNDr. Lucii Mazurové, Ph.D. za trpělivost, cenné rady a poskytnuté materiály.

Obsah

Úvod **3**

1 Úmrtnost **5**
1.1 Základní pojmy z demografie 5
1.2 Úmrtnostní tabulka 6
1.3 Modelování úmrtnosti 8

2 Stochastické modely úmrtnosti v jedné populaci **11**
2.1 Lee-Carterův model 12
2.1.1 Odhad parametrů 13
2.1.2 Modelování a předpovídání časové složky 16
2.1.3 Rezidua 17
2.2 Renshaw-Habermanův model 18
2.2.1 Odhad parametrů 18
2.2.2 Modelování a předpovídání časové složky 19
2.2.3 Rezidua 19
2.3 Cairns-Blake-Dowdův model 20
2.3.1 Odhad parametrů 20
2.3.2 Modelování a předpovídání časové složky 20
2.3.3 Rezidua 21

3 Modelování vývoje úmrtnosti v korelovaných populacích **22**
3.1 Rozšíření Lee-Carterova modelu 22
3.1.1 Nezávislý model 22
3.1.2 Model se společným K_1 23
3.1.3 Kointegrovaný model 24
3.1.4 Model společných faktorů 25
3.1.5 Rozšířený model společných faktorů 25
3.1.6 Model se společným A_2 26
3.2 Renshaw-Habermanův model pro více populací 27
3.3 Zobecnění Cairns-Blake-Dowdova modelu 27
3.4 Odhad metodou maximální věrohodnosti 28
3.5 Projekce časových řad 29
3.5.1 Projekce efektu kalendářního roku 29
3.5.2 Projekce kohortního efektu 30

4 Aplikace modelů na reálná data **31**
4.1 Úmrtnost české populace 31
4.2 Úmrtnost slovenské populace 35
4.3 Společné modelování úmrtnosti české a slovenské populace 38
4.3.1 Lee-Carterův model 38
4.3.2 Renshaw-Habermanův model 41
4.3.3 Cairns-Blake-Dowdův model 43
4.4 Předpověď úmrtnosti české a slovenské populace 43
4.4.1 Rozšířený Lee-Carterův model společných faktorů 45

4.4.2 Renshaw-Habermanův model společné kohorty 49

Závěr **54**

Seznam použité literatury **55**

Seznam obrázků **57**

Seznam tabulek **58**

Úvod

Modelování vývoje úmrtnosti v čase, zejména ve vyšších věcích, hraje významnou roli v oblasti životního a důchodového pojištění. Postupné snižování míry úmrtnosti v jednotlivých věcích, jehož důsledkem je prodlužující se naděje dožití, která je jednou z hlavních příčin stárnutí populace, patří mezi velmi diskutovaná témata, a to nejen na úrovni pojišťoven, ale i států a nadnárodních organizací.

Pro modelování úmrtnosti bylo od konce 20. století popsáno několik stochastických modelů, které jsou dodnes v praxi velmi často využívané. Mezi nejpopulárnější modely patří Lee-Carterův, Cairns-Blake-Dowdův a Renshaw-Habermanův model, případně jejich modifikace. Tyto modely ve své původní podobě však slouží k modelování úmrtnosti v samostatných populacích a nezachycují případnou souvislost s úmrtností v jiné populaci.

Úmrtnost v blízkých populacích je ale ve skutečnosti ovlivněna mnoha společnými faktory. Těmito faktory mohou být například epidemické výskyty chorob na určitém území, pokroky v léčbě závažných onemocnění, válečné konflikty nebo další dopady politické, sociální či ekonomické situace v daném regionu, které mohou být stále významnější díky integraci států do nadnárodních celků a rozvoji v oblasti dopravy a komunikace. Tyto aspekty vedou k představě, že rozdíly v úmrtnosti v blízkých populacích by se neměly z dlouhodobého hlediska zvyšovat.

Využití modelů pro současné modelování úmrtnosti ve více populacích umožňuje přistoupit k úloze modelování úmrtnosti komplexněji a zohlednit i úmrtnostní zkušenosti jiné blízké populace. Jinou populací nemusí být nutně myšlena populace disjunktní, naopak mohou v praxi tyto modely nalézt uplatnění v pojišťovnictví při modelování úmrtnosti v pojistném kmeni se zohledněním vývoje úmrtnosti v celé populaci daného státu.

Tato diplomová práce si klade za cíl popsat navrhovaná rozšíření stochastických modelů úmrtnosti v samostatných populacích na modely úmrtnosti ve dvou či více korelovaných populacích a následně ilustrovat tyto postupy na mírách úmrtnosti v populacích České a Slovenské republiky.

Samotná práce je rozčleněna do čtyř kapitol. V první části práce jsou definovány základní pojmy z oblasti demografie, které se využívají při modelování úmrtnosti, dále je v této kapitole popsán způsob zveřejňování údajů o úmrtnosti a jsou vymezeny základní metody užívané k jejímu modelování.
Navazuje kapitola 2, ve které jsou blíže popsány nejužívanější stochastické přístupy k modelování úmrtnosti – a sice výše zmíněný Lee-Carterův, Renshaw-Habermanův a Cairns-Blake-Dowdův model. Pro jednotlivé modely jsou popsány metody odhadu jejich parametrů a modely pro projekci vývoje úmrtnosti v budoucích letech.
Třetí kapitola následně uvádí možná zobecnění těchto modelů pro modelování

úmrtnosti ve více populacích současně. Tedy modely, které zohledňují závislost mezi těmito populacemi a projektují vývoj věkově specifické míry úmrtnosti tak, aby rozdíly v míře úmrtnosti mezi populacemi nenarůstaly.
Teoreticky popsané poznatky jsou poté ilustrovány na úmrtnostech v České a Slovenské republice v kapitole 4, ve které jsou srovnány projekce pomocí samostatných modelů a modelů pro současné modelování úmrtnosti v obou populacích. Pro výpočty byl využit software Wolfram Mathematica a EViews.

1. Úmrtnost

Kapitola *Úmrtnost* se zabývá zavedením některých základních pojmů z oblasti demografie, které jsou využívané v dalších kapitolách, dále popisuje údaje zveřejňované v úmrtnostních tabulkách a uvádí základní přístupy k modelování úmrtnosti. Získané poznatky pocházejí zejména z literatury [1] a [2], není-li uvedeno jinak.

1.1 Základní pojmy z demografie

Označme $T_x, x \geq 0$, náhodnou veličinu reprezentující *zbývající délku života* jedince ve věku x za podmínky, že se tohoto věku dožil. Předpokládejme, že T_x je spojitá náhodná veličina na $\mathbb{R}_+^0$, tedy zbývající délka života nabývá nezáporných reálných hodnot, a označme $f_x(t)$ její *hustotu* a $F_x(t)$ *(kumulativní) distribuční funkci.*

Hodnotu $F_x(t)$ je tedy možné interpretovat jako *pravděpodobnost úmrtí x-leté osoby před dožitím věku x+t*, pro její označení se standardně užívá symbol ${}_tq_x$ (pro $t = 1$ používejme značení q_x)

$$ {}_tq_x = F_x(t) = \mathsf{P}(T_x \leq t). $$

Často se vedle distribuční funkce zbývající délky života užívá také *funkce přežití*

$$ S_x(t) = 1 - F_x(t) = \mathsf{P}(T_x > t), $$

hodnota $S_x(t)$ poté značí pravděpodobnost, že jedinec ve věku x se dožije věku $x + t$, označme ji ${}_tp_x$

$$ {}_tp_x = S_x(t) = \mathsf{P}(T_x > t). $$

O zbývající délce života T_x předpokládáme, že

$$ \mathsf{P}(T_x \leq t) = \mathsf{P}(T_0 \leq x + t \mid T_0 > x), $$

tj. pro $F_x(t)$, $f_x(t)$ a $S_x(t)$ platí

$$ \begin{aligned} F_x(t) &= \mathsf{P}(T_x \leq t) = \mathsf{P}(T_0 \leq x + t \mid T_0 > x) = \frac{\mathsf{P}(x < T_0 \leq x + t)}{\mathsf{P}(T_0 > x)} = \\ &= \frac{F_0(x+t) - F_0(x)}{1 - F_0(x)} = \frac{F_0(x+t) - F_0(x)}{S_0(x)}, \qquad (1.1) \\ f_x(t) &= \frac{\partial}{\partial t} F_x(t) = \frac{\frac{\partial}{\partial t}\left(F_0(x+t) - F_0(x)\right)}{S_0(x)} = \frac{f_0(x+t)}{S_0(x)}, \qquad (1.2) \\ S_x(t) &= \mathsf{P}(T_x > t) = \mathsf{P}(T_0 > x + t \mid T_0 > x) = \frac{\mathsf{P}(T_0 > x + t)}{\mathsf{P}(T_0 > x)} = \frac{S_0(x+t)}{S_0(x)}. \end{aligned} $$

Střední hodnota náhodné veličiny T_x se v demografii označuje jako *střední délka života jedince ve věku x*

$$e_x = \mathsf{E}\left(T_x\right) = \int_0^\infty t \cdot f_x(t)dt.$$

Intenzita úmrtnosti jedince ve věku x je, dle [3], označovaná symbolem μ_x a definovaná vztahem

$$\mu_x = \lim_{t\to\ 0} \frac{\mathsf{P}(T_x \leq t)}{t} = \lim_{t\to\ 0} \frac{{}_tq_x}{t}.$$

S využitím vztahu 1.1 můžeme μ_x přepsat do tvaru

$$\begin{aligned}\mu_x &= \lim_{t\to\ 0} \frac{F_0(x+t) - F_0(x)}{t \cdot (1 - F_0(x))} = \lim_{t\to\ 0} \frac{F_0(x+t) - F_0(x)}{t \cdot S_0(x)} = \\ &= \frac{\partial F_0(x)}{\partial x} \cdot \frac{1}{S_0(x)} = \frac{f_0(x)}{S_0(x)},\end{aligned} \tag{1.3}$$

respektive

$$\begin{aligned}\mu_x &= \lim_{t\to\ 0} \frac{F_0(x+t) - F_0(x)}{t \cdot (1 - F_0(x))} = \frac{\partial F_0(x)}{\partial x} \cdot \frac{1}{1 - F_0(x)} = \frac{-\frac{\partial(1-F_0(x))}{\partial x}}{1 - F_0(x)} = \\ &= -\frac{\partial}{\partial x} \ln(1 - F_0(x)) = -\frac{\partial}{\partial x} \ln({}_xp_0).\end{aligned}$$

Pomocí vztahů 1.2 a 1.3 lze dále snadno získat

$$\mu_{x+t} = \frac{f_0(x+t)}{S_0(x+t)} = \frac{f_x(t)}{1 - F_x(t)} = -\frac{\partial}{\partial t} \ln(1 - F_x(t)) = -\frac{\partial}{\partial t} \ln_t p_x.$$

Alternativně lze tedy pomocí intenzity úmrtnosti vyjádřit pravděpodobnost dožití ${}_tp_x$ vztahem

$${}_tp_x = \exp\left(-\int_x^{x+t} \mu_u du\right) = \exp\left(-\int_0^t \mu_{x+u} du\right).$$

1.2 Úmrtnostní tabulka

Úmrtnostní tabulky představují model vymírání teoretické populace jedinců s počátečním rozsahem l_0. Dle [3] můžeme vyčlenit dva druhy úmrtnostních tabulek:

- *generační (kohortní)* – úmrtnostní tabulky založené na sledování aktuálního počtu živých jedinců ve věku $1, 2, ..., \omega$, pocházejících ze skupiny l_0 jedinců, kteří se narodili stejném kalendářním roce,

- *průřezové (běžné)* – úmrtnostní tabulky, které se v daném kalendářním roce zakládají na úmrtnosti jedinců, kteří se právě v tomto roce nacházejí ve věku $1, 2, ..., \omega$,

kde ω označuje *maximální věk dožití* – Český statistický úřad například uvažuje $\omega = 105$. Konstrukce generačních úmrtnostních tabulek tedy zřejmě vyžaduje sledování období dlouhého $\omega + 1$ let.

Jednotlivé ukazatele uváděné v úmrtnostních tabulkách jsou konstruovány na základě pravděpodobností úmrtí q_x a pravděpodobností přežití p_x, které byly definovány v podkapitole 1.1. Jedná se zejména o funkce

- l_x – střední počet jedinců, kteří se při daném výchozím stavu l_0 (nejčastěji se uvažuje hodnota $l_0 = 100000$) dožijí věku x, hodnotu l_0 nazýváme kořenem úmrtnostní tabulky,

  $$l_x = l_0 \cdot_x p_0, \qquad \text{resp. } l_{x+n} = l_x \cdot_n p_x.$$

  Mezi víceletými pravděpodobnostmi dožití/úmrtí a počty dožívajících se platí vztahy

  $${}_np_x = \frac{l_{x+n}}{l_x}, \qquad {}_nq_x = \frac{l_x - l_{x+n}}{l_x}.$$

- d_x – střední počet jedinců, kteří při výchozím stavu l_0 zemřou ve věku mezi x a $x + 1$,

  $$d_x = l_x - l_{x+1}.$$

Úmrtnostní tabulky obvykle dále uvádějí také hodnoty zavedené pro účely aproximace střední délky života:

- L_x – střední počet "člověkoroků", které ve věku x celkem prožije l_x jedinců

  $$L_x = l_{x+1} + \frac{1}{2}d_x = \frac{l_x + l_{x+1}}{2}.$$

- $\mathbb{T}_x$ – (střední) celkový počet zbývajících let života jedinců, kteří jsou naživu ve věku x

  $$\mathbb{T}_x = L_x + L_{x+1} + L_{x+2} + ... + L_\omega.$$

- $\overline{e}_x$ – naděje dožití (střední délka života ve věku x)

  $$\overline{e}_x = \frac{\mathbb{T}_x}{l_x}.$$

Jelikož z úmrtnostní tabulky známe zpravidla údaje o úmrtnosti pouze pro celočíselné věky, zatímco náhodnou veličinu T_x reprezentující zbývající dobu života x-leté osoby uvažujeme spojitou, je v demografii zvykem užívat následující předpoklady:

- *předpoklad področní linearity úmrtnosti*: pro $t \in \langle 0; 1\rangle$ platí ${}_tq_x = t \cdot q_x$,

- *předpoklad konstantní intenzity úmrtnosti*: pro $t \in (0; 1)$ platí $\mu_{x+t} = \mu$, kde x je pevně zvolený celočíselný věk a konstantu μ získáme ze vztahu

$$\mu = -\ln p_x,$$

- *Balducciho předpoklad*: pro $t \in \langle 0; 1\rangle$ platí ${}_{1-t}q_{x+t} = (1-t) \cdot q_x$.

Nezbytným krokem pro výpočet výše uvedených údajů úmrtnostní tabulky je odhad pravděpodobnosti úmrtí. Dle metodiky Českého statistického úřadu [4] je možné ji za předpokladu konstantní intenzity úmrtnosti odhadnout jako

$$q_x = 1 - \exp(-m_x), \tag{1.4}$$

kde m_x značí *věkově specifickou míru úmrtnosti*, která je definovaná vztahem

$$m_x = \frac{S_0(x) - S_0(x+1)}{\int_0^1 S_0(x+u)du}$$

a pro reálnou populaci je možné ji určit pomocí vzorce

$$m_x = \frac{D_x}{E_x}, \quad x = 1, 2, 3, ...,$$

ve kterém D_x označuje absolutní počet zemřelých osob ve věku x na zkoumaném území v daném roce a E_x absolutní počet obyvatel ve věku x k 1. červenci daného roku na zkoumaném území. Získané hodnoty q_x se v praxi ještě dále upravují – vyrovnávají pomocí klouzavých průměrů nebo extrapolují pomocí Gompertzova-Makehamova vzorce.

1.3 Modelování úmrtnosti

Jak uvádí [3], ukazuje se, že úmrtnost v jednotlivých věcích se v čase mění – věk, ve kterém dochází k největšímu počtu úmrtí se s časem posouvá k vyšším věkům (tento efekt nazýváme *expanzí*) a zároveň se okolo něj úmrtí v ostatních věcích čím dál více koncentrují (dochází k tzv. *rektangularizaci* křivky přežívání zachycující hodnoty l_x z úmrtnostní tabulky). Z tohoto důvodu se pro modelování úmrtnosti využívají dynamické modely, tedy modely, které předpokládají, že pravděpodobnosti úmrtí (případně další sledované ukazatele) nezávisí pouze na věku osoby x, ale jsou i funkcí roku, ve kterém jedinec daného věku dosáhne. Například namísto jednoleté pravděpodobnosti úmrtí q_x tedy v dynamickém modelu uvažujeme $q_x(t)$, funkci věku a kalendářního roku.

Při dynamickém modelování úmrtnosti tedy pravděpodobnosti úmrtí z průřezových úmrtnostních tabulek z po sobě jdoucích kalendářních let tvoří matici, jejíž řádky odpovídají věkům a sloupce rokům. Uvažujeme-li celočíselné věky, můžeme například matici pravděpodobností úmrtí ve věku x zapsat ve tvaru

$$\begin{pmatrix} & \vdots & \vdots & \vdots & \\ \dots & q_{x-1}(t-1) & q_{x-1}(t) & q_{x-1}(t+1) & \dots \\ \dots & q_x(t-1) & q_x(t) & q_x(t+1) & \dots \\ \dots & q_{x+1}(t-1) & q_{x+1}(t) & q_{x+1}(t+1) & \dots \\ & \vdots & \vdots & \vdots & \end{pmatrix}.$$

Na prvky uvedené matice poté můžeme nahlížet

- *vertikálně* – $q_0(t)$, $q_1(t)$, $q_2(t)$, ... – jako na posloupnost pravděpodobností úmrtí uvedenou v úplné průřezové úmrtnostní tabulce,

- *diagonálně* – $q_0(t)$, $q_1(t+1)$, $q_2(t+2)$, ... – jako na posloupnost pravděpodobností úmrtí uvedenou v kohortní úmrtnostní tabulce pro kohortu narozenou v roce t,

- *horizontálně* – $q_x(t)$, $q_x(t+1)$, $q_x(t+2)$, ... – jako na *úmrtnostní profil* vztahující se k věku x.

Zaměřme se dále na extrapolační metody, které předpokládají, že trend pozorovaný v minulých datech bude přetrvávat i do budoucna. Jejich cílem je popsat úmrtnost matematickou funkcí, jejímž grafem je hladká křivka. Přestože se extrapolační metody hodí spíše pro krátkodobé předpovídání (jelikož nezohledňují limity lidského těla a další, například medicínské, faktory), jsou v aktuárské praxi nejrozšířenější. Podle přístupu k pozorovaným hodnotám můžeme rozdělit modely úmrtnosti na:

- *deterministické*, které předpokládají, že hodnoty uvedené v úmrtnostních tabulkách jsou pevně daná čísla,

- *stochastické*, u kterých považujeme hodnoty uvedené v úmrtnostních tabulkách za realizace náhodných veličin.

Při deterministickém modelování úmrtnosti, konkrétně modelování pravděpodobnosti úmrtí ve věku x, se při horizontálním přístupu k datům často využívá předpoklad, že vývoj logaritmu pravděpodobnosti úmrtí ve věku x je pro všechna x lineární. Tedy, že pro $h = 1, 2, ..., n-1$ platí

$$\log q_x(t_{h+1}) - \log q_x(t_h) \approx -\delta_x(t_{h+1} - t_h),$$

neboli

$$\frac{q_x(t_{h+1})}{q_x(t_h)} \approx \exp\{-\delta_x(t_{h+1} - t_h)\},$$

kde pro každý věk x je parametr δ_x možné odhadnout například pomocí metody nejmenších čtverců. Tento přístup nazýváme *exponenciální formulí.*

Naopak při vertikálním přístupu k datům, modelování úmrtnosti v daném roce t, se při deterministickém modelování využívá *úmrtnostních zákonů.* Jednotlivé zákony jsou charakterizovány typickým tvarem funkce popisující intenzity úmrtnosti nebo pravděpodobnosti úmrtí. Mezi nejznámější úmrtnostní zákony patří:

- *Gompertzův zákon* (1825): $\mu_x(t) = B(t) \cdot C(t)^x$,

- *Gompertzův-Makehamův zákon* (1890): $\mu_x(t) = A(t) + B(t) \cdot C(t)^x$,

- *(Druhý) Heligmanův-Pollardův zákon* (1980):

$$q_x(t) = A(t)^{(x+B(t))^{C(t)}} + D(t) \cdot e^{-E(t)(\log x - \log F(t))^2} + \frac{G(t) \cdot H(t)^x}{1 + G(t) \cdot H(t)^x},$$

kde jednotlivé parametry $A(t)$, $B(t)$, $C(t)$, $D(t)$, $E(t)$, $F(t)$, $G(t)$ a $H(t)$ můžeme opět odhadnout pomocí metody nejmenších čtverců.

2. Stochastické modely úmrtnosti v jedné populaci

Dle [3] považujeme při stochastickém modelování úmrtnosti pozorované míry úmrtnosti za realizace náhodných veličin reprezentujících minulou úmrtnost. Projektované míry úmrtnosti poté chápeme jako odhady náhodných veličin reprezentujících budoucí úmrtnost. Stochastický přístup nám umožňuje určit mimo bodových odhadů úmrtnosti i odhady intervalové, na druhou stranu však vyžaduje splnění jistých statistických předpokladů (rozdělení počtu úmrtí).

Pro stochastické modelování úmrtnosti bylo v minulosti vytvořeno několik modelů. Za první z nich a zároveň v současnosti stále nejpoužívanější je možné považovat Lee-Carterův model z roku 1992. Tento model byl několikrát modifikován, jednou z používaných modifikací je Renshaw-Habermanův model. Dalším ze základních typů modelů úmrtnosti je Cairns-Blake-Dowdův model (též označovaný zkratkou jmen autorů – CBD), k němuž bylo rovněž odvozeno několik modifikací, jimi se však nebudeme podrobněji zabývat.

Některé modely modelují logaritmus věkově specifické míry úmrtnosti $m_x(t)$, jiné logitovou transformaci pravděpodobnosti úmrtí osoby ve věku x, označovanou $\text{logit}(q_x(t))$. Míru úmrtnosti lze za předpokladu konstantní intenzity úmrtnosti vyjádřit s využitím vztahu 1.4 jako

$$m_x(t) = -\log(1 - q_x(t)).$$

V případě modelu CBD (a jeho modifikací) je možné využít zmíněnou logitovou transformaci, připomeňme proto, že platí

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \log(p) - \log(1-p).$$

Přehled nejpoužívanějších modelů dle [5] je uvedený v tabulce 2.1. V jednotlivých modelech zastupují

$a_1(x), a_2(x), a_3(x)$ funkce popisující efekt věku x,

$k_1(t), k_2(t)$ funkce popisující efekt kalendářního roku t,

$c(t-x)$ funkce vztahující se ke kohortnímu efektu, tedy funkce zahrnující vliv roku narození $t-x$.

V tabulce je možné si povšimnout, že některé modely nevyužívají složku kohortního efektu, takové modely obvykle nazýváme *age-period modely*. Naopak v případě zahrnutí kohortního efektu se jedná o takzvané *age-period-cohort modely* (také se užívá zkratka *APC*).

Následující podkapitoly se podrobněji věnují Lee-Carterovu, Renshaw-Habermanovu a Cairns-Blake-Dowdovu modelu. Uvedené modely jsou specifikovány, je pro ně shrnuta metodika odhadu parametrů a dále popsána tvorba (bodových)

Název	Formulace
Lee-Carter	$\log(m_x(t)) = a_1(x) + a_2(x) \cdot k_1(t)$
Renshaw-Haberman	$\log(m_x(t)) = a_1(x) + a_2(x) \cdot k_1(t) + a_3(x) \cdot c(t-x)$
APC model (Currie)	$\log(m_x(t)) = a_1(x) + m^{-1} \cdot k_1(t) + (n_a)^{-1} \cdot c(t-x)$
Cairns-Blade-Dowd	$\mathrm{logit}(q_x(t)) = k_1(t) + x \cdot k_2(t)$
Zobecněný CBD	$\mathrm{logit}(q_x(t)) = k_1(t) + x \cdot k_2(t) + c(t-x)$

Pozn: m označuje počet věků zahrnutých v datovém souboru

Tabulka 2.1: Přehled nejpoužívanějších stochastických modelů

projekcí. *Intervalové* odhady však nejsou popsány analyticky, neboť je složité při předpovědi zahrnout různé zdroje nejistoty (nepřesnost v odhadu parametrů, nejistota v projekci časové složky do budoucnosti). Proto se při vytváření intervalových odhadů budoucího vývoje úmrtnosti v praxi využívá metoda *bootstrap*.

2.1 Lee-Carterův model

Jak navrhují R. D. Lee a L. R. Carter ve své práci [6] z roku 1992, uvažujme model pro logaritmus věkově specifické míry úmrtnosti, který je možné zapsat ve tvaru

$$\log(m_x(t)) = a_1(x) + a_2(x) \cdot k_1(t) + \varepsilon_{x,t}, \tag{2.1}$$

kde $a_1(x)$ a $a_2(x)$ zastupují vhodně zvolené věkově specifické konstanty a $k_1(t)$ označuje s časem se měnící index, tento index je možné interpretovat jako úroveň úmrtnosti v roce t.
Hodnota $a_1(x)$ tedy udává průměrnou závislost úmrtnosti na věku a hodnota $a_2(x)$ říká, jak rychle klesá s časem úmrtnost v jednotlivých věcích oproti ostatním věkům. Teoreticky by pro nějaký věk x mohlo být i $a_2(x) < 0$, což by znamenalo, že ve věku x míra úmrtnosti s časem roste, zatímco v ostatních věcích klesá, což v praxi (zejména v dlouhodobém horizontu) nenastává.
$\varepsilon_{x,t}$ značí náhodné odchylky s nulovou střední hodnotou a rozptylem σ_ε^2. Splnění předpokladu homoskedasticity je v praxi diskutabilní, neboť variabilita logaritmu skutečné věkově specifické míry úmrtnosti je ve vyšších věcích větší než u mladších jedinců. Ukazuje se však, že pokud omezíme rozsah pro modelování například na úmrtnost ve věcích vyšších než 50 let, čímž se vyhneme nestabilní oblasti úmrtnosti 20–30-letých jedinců (takzvaný *accident hump*), nečiní tento předpoklad dále potíže.

Předpokládejme, že známe věkově specifické míry úmrtnosti jedinců ve věcích $x_1, \ldots, x_m$ v po sobě jdoucích kalendářních letech $t_1, \ldots, t_n$. Parametry modelu nejsou určeny jednoznačně, proto se nejčastěji zavádí omezující podmínky

$$\sum_{x=x_1}^{x_m} a_2(x) = 1 \quad \text{a} \quad \sum_{t=t_1}^{t_n} k_1(t) = 0. \tag{2.2}$$

2.1.1 Odhad parametrů

K odhadu parametrů $a_1(x_1), \ldots, a_1(x_m)$, $a_2(x_1), \ldots, a_2(x_m), k_1(t_1), \ldots, k_1(t_n)$ Lee-Carterova modelu na základě pozorovaných hodnot $m_{x_1}(t_1), \ldots, m_{x_m}(t_n)$ se dle [3] využívají nejčastěji metody:

- Metoda nejmenších čtverců – singulární rozklad,
- Metoda nejmenších čtverců – Newton-Raphsonův algoritmus,
- Poissonovský model.

Všimněme si, že nejde o klasický regresní model, neboť na pravé straně rovnice 2.1 se nenachází žádná pozorovaná hodnota, ale pouze parametry modelu, které chceme odhadnout.

Metoda nejmenších čtverců – singulární rozklad

Úlohou metody nejmenších čtverců (v případě využití singulárního rozkladu i Newton-Raphsonova iteračního algoritmu) je najít odhady $\hat{a}_1(x_1), \ldots, \hat{a}_1(x_m)$, $\hat{a}_2(x_1), \ldots, \hat{a}_2(x_m)$ a $\hat{k}_1(t_1), \ldots, \hat{k}_1(t_n)$ minimalizující

$$O_{LS}(\mathbf{a_1}, \mathbf{a_2}, \mathbf{k_1}) = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} \left(\log(m_x(t)) - a_1(x) - a_2(x) \cdot k_1(t)\right)^2. \tag{2.3}$$

Položením parciální derivace O_{LS} podle $a_1(x)$, $x = x_1, \ldots, x_m$ rovné nule získáme vztah

$$\sum_{t=t_1}^{t_n} \log(m_x(t)) = (t_n - t_1 + 1) \cdot \hat{a}_1(x) + \hat{a}_2(x) \sum_{t=t_1}^{t_n} \hat{k}_1(t),$$

odkud je možné s využitím omezující podmínky 2.2 pro $\hat{k}_1(t)$ vyjádřit odhad pro $a_1(x)$ jako

$$\hat{a}_1(x) = \frac{1}{(t_n - t_1 + 1)} \sum_{t=t_1}^{t_n} \log(m_x(t)), \quad x = x_1, \ldots, x_m.$$

Odhady parametrů $a_1(x)$ jsou tedy pro každé x rovné aritmetickému průměru $\log(m_x(t))$ přes index kalendářního roku.

Pro odhad zbylých parametrů modelu $(a_2(x_1), \ldots, a_2(x_m), k_1(t_1), \ldots, k_1(t_n))$ využijme singulárního rozkladu matice $\mathbf{Z}$ s prvky $z_{x,t} = \log m_x(t) - \hat{a}_1(x)$ dimenze $(x_m - x_1 + 1) \times (t_n - t_1 + 1)$. Hledejme tedy odhady minimalizující

$$O_{LS}^{SVD}(\mathbf{a_2}, \mathbf{k_1}) = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} \left(z_{x,t} - a_2(x) \cdot k_1(t)\right)^2.$$

Označme λ_1 největší vlastní číslo matice $\mathbf{Z}^\top\mathbf{Z}$, $\mathbf{u_1}$ vlastní vektor odpovídající vlastnímu číslu λ_1 matice $\mathbf{Z}^\top\mathbf{Z}$ a obdobně $\mathbf{v_1}$ příslušný vlastní vektor matice $\mathbf{Z}\mathbf{Z}^\top$. Nejlepší aproximaci $\mathbf{Z}$ ve smyslu nejmenších čtverců založenou na vektorech $\mathbf{u_1}$ a $\mathbf{v_1}$ a hodnotě λ_1 potom můžeme zapsat jako

$$\mathbf{Z} \approx \sqrt{\lambda_1}\mathbf{v_1}\mathbf{u_1}^\top,$$

odkud lze odvodit, že

$$\hat{\mathbf{a}}_{\mathbf{2}} = \frac{\mathbf{v_1}}{\sum_{j=1}^{x_m-x_1+1} v_{1j}} \qquad \text{a} \qquad \hat{\mathbf{k}}_{\mathbf{1}} = \sqrt{\lambda_1}\left(\sum_{j=1}^{x_m-x_1+1} v_{1j}\right)\mathbf{u_1},$$

pro $\sum_{j=1}^{x_m-x_1+1} v_{1j} \neq 0$.

Metoda nejmenších čtverců – Newton-Raphsonův algoritmus

Alternativní metodou k řešení úlohy minimalizace nejmenších čtverců 2.3 je využití Newton-Raphsonova iteračního algoritmu. Funkce $O_{LS}(\mathbf{a_1}, \mathbf{a_2}, \mathbf{k_1})$ nabývá minimální hodnoty, pokud každá z parciálních derivací (podle $a_1(x), a_2(x)$ a $k_1(t)$) je rovna nule, tj. pokud

$$\begin{aligned}
0 &= \sum_{t=t_1}^{t_n} \Big(\log(m_x(t)) - \hat{a}_1(x) - \hat{a}_2(x)\cdot\hat{k}_1(t)\Big), && x = x_1, \ldots, x_m,\\
0 &= \sum_{x=x_1}^{x_m} \hat{a}_2(x)\Big(\log(m_x(t)) - \hat{a}_1(x) - \hat{a}_2(x)\cdot\hat{k}_1(t)\Big), && t = t_1, \ldots, t_n, \qquad (2.4)\\
0 &= \sum_{t=t_1}^{t_n} \hat{k}_1(t)\Big(\log(m_x(t)) - \hat{a}_1(x) - \hat{a}_2(x)\cdot\hat{k}_1(t)\Big), && x = x_1, \ldots, x_m.
\end{aligned}$$

Newton-Raphsonova iterační formule pro řešení obecné rovnice $f(y) = 0$ má tvar

$$y^{(k+1)} = y^{(k)} - \frac{f(y^{(k)})}{f'(y^{(k)})},$$

kde za $y^{(0)}$ zvolíme vhodnou počáteční hodnotu. Pro odhadnutí parametrů $a_1(x)$, $a_2(x)$ a $k_1(t)$, $x = x_1, \ldots, x_m$, $t = t_1, \ldots, t_n$ ze soustavy rovnic 2.4 navrhuje [3] postupně použít kroky

$$\begin{aligned}
\hat{a}_1^{(k+1)}(x) &= \hat{a}_1^{(k)}(x) + \frac{\sum_{t=t_1}^{t_n}\Big(\log(m_x(t)) - \hat{a}_1^{(k)}(x) - \hat{a}_2^{(k)}(x)\cdot\hat{k}_1^{(k)}(t)\Big)}{t_n - t_1 + 1},\\
\hat{k}_1^{(k+1)}(t) &= \hat{k}_1^{(k)}(t) + \frac{\sum_{x=x_1}^{x_m}\hat{a}_2^{(k)}(x)\Big(\log(m_x(t)) - \hat{a}_1^{(k+1)}(x) - \hat{a}_2^{(k)}(x)\cdot\hat{k}_1^{(k)}(t)\Big)}{\sum_{x=x_1}^{x_m}\Big(\hat{a}_2^{(k)}(x)\Big)^2},\\
\hat{a}_2^{(k+1)}(x) &= \hat{a}_2^{(k)}(x) +\\
&\quad + \frac{\sum_{t=t_1}^{t_n}\hat{k}_1^{(k+1)}(t)\Big(\log(m_x(t)) - \hat{a}_1^{(k+1)}(x) - \hat{a}_2^{(k)}(x)\cdot\hat{k}_1^{(k+1)}(t)\Big)}{\sum_{t=t_1}^{t_n}\Big(\hat{k}_1^{(k+1)}(t)\Big)^2},
\end{aligned}$$

za inicializační hodnoty algoritmu se dle [7] ukazuje vhodné použít

$$\hat{a}_1^{(0)}(x) = (t_n - t_1 + 1)^{-1}\sum_{t=t_1}^{t_n}\log(m_x(t)), \quad \hat{k}_1^{(0)}(t) = 0 \quad \text{a} \quad \hat{a}_2^{(0)}(x) = m^{-1}.$$

Poissonovský model

Poissonovský model umožňuje odhad vývoje úmrtnosti pomocí metody maximální věrohodnosti. Narozdíl od výše uvedených metod nepracuje s věkově specifickými mírami úmrtnosti, ale s počty úmrtí osob ve věku x v roce t, ozn. $D_x(t)$, a s příslušnými expozicemi $E_x(t)$ (např. počet jedinců ve věku x v polovině roku t). O počtu úmrtí ve věku x v roce t předpokládáme, že se jedná o náhodnou veličinu s Poissonovým rozdělením

$$D_x(t) \sim Po\left(E_x(t) \cdot \exp\{a_1(x) + a_2(x) \cdot k_1(t)\}\right).$$

Logaritmickou věrohodnostní funkci poté můžeme za předpokladu nezávislosti počtu úmrtí zapsat jako

$$\ell(\mathbf{a_1}, \mathbf{a_2}, \mathbf{k_1}) = \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} [D_x(t)(a_1(x) + a_2(x) \cdot k_1(t)) - \\ - E_x(t) \cdot \exp\{a_1(x) + a_2(x) \cdot k_1(t)\}] + c,$$

kde c je konstanta, která nezávisí na odhadovaných parametrech. Při maximalizaci $\ell(\mathbf{a_1}, \mathbf{a_2}, \mathbf{k_1})$ položíme parciální derivace podle parametrů $a_1(x), a_2(x), k_1(t)$ rovné nule a dále můžeme najít řešení opět pomocí Newton-Raphsonova iteračního algoritmu, jednotlivé kroky lze dle [3] zapsat pomocí formulí:

$$\hat{a}_1^{(k+1)}(x) = \hat{a}_1^{(k)}(x) + \frac{\sum_{t=t_1}^{t_n} \left(D_x(t) - E_x(t) \cdot \exp\{\hat{a}_1^{(k)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k)}(t)\}\right)}{\sum_{t=t_1}^{t_n} E_x(t) \cdot \exp\{\hat{a}_1^{(k)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k)}(t)\}},$$

$$\hat{k}_1^{(k+1)}(t) = \hat{k}_1^{(k)}(t) + \\ + \frac{\sum_{x=x_1}^{x_m} \left(D_x(t) - E_x(t) \cdot \exp\{\hat{a}_1^{(k+1)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k)}(t)\}\right) \hat{a}_2^{(k)}(x)}{\sum_{x=x_1}^{x_m} E_x(t) \cdot \exp\{\hat{a}_1^{(k+1)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k)}(t)\} \left(\hat{a}_2^{(k)}(x)\right)^2},$$

$$\hat{a}_2^{(k+1)}(x) = \hat{a}_2^{(k)}(x) + \\ + \frac{\sum_{t=t_1}^{t_n} \left(D_x(t) - E_x(t) \cdot \exp\{\hat{a}_1^{(k+1)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k+1)}(t)\}\right) \hat{k}_1^{(k+1)}(t)}{\sum_{t=t_1}^{t_n} E_x(t) \cdot \exp\{\hat{a}_1^{(k+1)}(x) + \hat{a}_2^{(k)}(x) \cdot \hat{k}_1^{(k+1)}(t)\} \left(\hat{k}_1^{(k+1)}(t)\right)^2},$$

$x = x_1, \ldots, x_m, t = t_1, \ldots, t_n$. Za počáteční hodnoty je vhodné zvolit $\hat{a}_1^{(0)}(x) = 0$, $\hat{a}_2^{(0)}(x) = 1$ a $\hat{k}_1^{(0)}(t) = 0$.

Korekce odhadů

Řadu odhadů $a_2(x), x = x_1, \ldots, x_m$, je třeba před použitím k projekci úmrtnosti vyhladit, jednou z možností, jak zajistit její hladký průběh, je například při použití metody nejmenších čtverců minimalizovat výraz

$$\sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} \left(\log(m_x(t)) - a_1(x) - a_2(x) \cdot k_1(t)\right)^2 + \\ + p \cdot \sum_{x=x_1}^{x_m} \left(a_2(x+2) - 2a_2(x+1) + a_2(x)\right)^2$$

namísto původně uvažované funkce O_{LS}. Parametr p poté nazýváme *vyhlazovacím parametrem.*

Jak navrhuje původní článek [6], je dále vhodné po získání odhadů $\hat{a}_1(x)$, $\hat{a}_2(x)$, $x = x_1, \ldots, x_m$, metodou nejmenších čtverců ještě jednou přepočítat $\hat{k}_1(t)$ tak, aby odhadovaná úmrtnost co nejlépe odpovídala pozorované úmrtnosti v čase t, $t = t_1, \ldots, t_n$, tj. aby pro každé t platilo

$$\sum_{x=x_1}^{x_m} D_x(t) = \sum_{x=x_1}^{x_m} E_x(t) \cdot \exp\{\hat{a}_1(x) + \hat{a}_2(x) \cdot \hat{k}_1(t)\}. \tag{2.5}$$

V případě odhadu parametrů Lee-Carterova modelu pomocí metody nejmenších čtverců i pomocí poissonovského modelu se výsledné odhady upravují tak, aby splňovaly stanovené omezující podmínky 2.2:

$$\begin{aligned} \hat{a}_1(x) &\rightarrow \hat{a}_1(x) + \hat{a}_2(x) \cdot \bar{k}_1, \\ \hat{k}_1(t) &\rightarrow (\hat{k}_1(t) - \bar{k}_1)\hat{a}_{2\bullet}, \\ \hat{a}_2(x) &\rightarrow \frac{\hat{a}_2(x)}{\hat{a}_{2\bullet}}, \end{aligned}$$

kde

$$\begin{aligned} \bar{k}_1 &= \tfrac{1}{n} \textstyle\sum_{t=t_1}^{t_n} \hat{k}_1(t), \\ \hat{a}_{2\bullet} &= \textstyle\sum_{x=x_1}^{x_m} \hat{a}_2(x). \end{aligned}$$

2.1.2 Modelování a předpovídání časové složky

Na $k_1(t)$ je dle [6] možné pohlížet jako na realizaci časové řady řídící se autoregresním integrovaným procesem klouzavých průměrů $ARIMA(p, d, q)$:

$$\begin{aligned} \Delta^d k_1(t) &= \mu + \phi_1 \Delta^d k_1(t-1) + \phi_2 \Delta^d k_1(t-2) + \cdots + \phi_p \Delta^d k_1(t-p) \\ &+ \xi_t + \psi_1 \xi_{t-1} + \psi_2 \xi_{t-2} + \cdots + \psi_q \xi_{t-q}, \end{aligned}$$

kde $\phi_p \neq 0$, $\psi_q \neq 0$, $\Delta^d k_1(t)$ značí d-tou diferenci procesu $\{k_1(t)\}$ a $\{\xi_t\}$ je Gaussovský bílý šum s rozptylem $\sigma_\xi^2 > 0$. V praxi se ukazuje, že často je vhodným modelem pro $k_1(t)$ náhodná procházka s driftem, tedy proces $ARIMA(0, 1, 0)$

$$k_1(t) = k_1(t-1) + d + \xi_t,$$

kde $\{\xi_t\}$ je opět Gaussovský bílý šum s rozptylem $\sigma^2 > 0$ a d je parametr driftu.

Projekci pro $t' > t_n$ v případě modelování $k_1(t)$ jako náhodné procházky s driftem je možné zapsat jako

$$\hat{m}_x(t') = \exp\{\hat{a}_1(x) + \hat{a}_2(x) \cdot \tilde{k}_1(t')\},$$

kde $\tilde{k}_1(t')$ značí bodový odhad časové složky a platí pro něj

$$\begin{aligned} \tilde{k}_1(t') &= \mathsf{E}\,(k_1(t') \mid k_1(t_1), \ldots, k_1(t_n)) = \\ &= \mathsf{E}\left(k_1(t_n) + (t' - t_n) \cdot d + \sum_{j=1}^{t'-t_n} \xi_{t_n+j} \mid k_1(t_1), \ldots, k_1(t_n)\right) = \\ &= k_1(t_n) + d \cdot (t' - t_n). \end{aligned}$$

Pro podmíněný rozptyl $k_1(t')$ platí

$$\operatorname{var}\left(k_1(t') \mid k_1(t_1), \ldots, k_1(t_n)\right) = (t' - t_n) \cdot \sigma^2.$$

Druhou možností je projektovat úmrtnost na základě poslední známé hodnoty míry úmrtnosti $m_x(t_n)$, v takovém případě můžeme bodový odhad $\hat{m}_x(t')$ vyjádřit jako

$$\hat{m}_x(t') = m_x(t_n) \cdot \exp\{\hat{a}_2(x) \cdot (\tilde{k}_1(t') - \hat{k}_1(t_n))\}.$$

Použití náhodné procházky s driftem znamená, že pro každý věk x se snižuje očekávaná věkově specifická míra úmrtnosti exponenciálně rychlostí závislou na parametru $a_2(x)$:

$$\log \hat{m}_x(t+1) - \log \hat{m}_x(t) = \hat{a}_2(x) \cdot \mathsf{E}\left(k_1(t+1) - k_1(t)\right) = \hat{a}_2(x) \cdot d\,.$$

Jelikož pro rozdíl $k_1(t) - k_1(t-1),\ t = t_2, \ldots, t_n$, platí, že se jedná o nezávislé normálně rozdělené náhodné veličiny se střední hodnotou d a rozptylem σ^2, je možné získat pomocí metody maximální věrohodnosti odhady parametrů d a σ^2 ze vzorců

$$\begin{aligned}
\hat{d} &= \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} (\hat{k}_1(t) - \hat{k}_1(t-1)) = \frac{\hat{k}_1(t_n) - \hat{k}_1(t_1)}{t_n - t_1}, \\
\hat{\sigma^2} &= \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} (\hat{k}_1(t) - \hat{k}_1(t-1) - \hat{d})^2.
\end{aligned}$$

2.1.3 Rezidua

Správnost modelu v případě, že byly parametry odhadnuty pomocí metody nejmenších čtverců, je možné dle [3] diagnostikovat pomocí grafického znázornění standardizovaných Pearsonových reziduí. Jednotlivá rezidua lze vypočítat ze vzorce

$$r_x(t) = \frac{\log(m_x(t)) - (\hat{a}_1(x) + \hat{a}_2(x) \cdot \hat{k}_1(t))}{\sqrt{\frac{1}{(x_m - x_1)(t_n - t_1 - 1)} \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} [\log(m_x(t)) - (\hat{a}_1(x) + \hat{a}_2(x) \cdot \hat{k}_1(t))]^2}}.$$

V případě poissonovského modelu je pro diagnostiku reziduí vhodné použít znázornění devianční reziduí definovaných vztahem

$$r_x(t) = sign(D_x(t) - \hat{D}_x(t)) \sqrt{dev(x,t)},$$

kde $dev(x,t) = 2\{D_x(t) \cdot \log(D_x(t)/\hat{D}_x(t)) - (D_x(t) - \hat{D}_x(t))\}$.

Pokud je z grafického znázornění zřejmé, že (v určité jeho části) rezidua nejsou náhodně rozprostřena, ale systematicky sledují nějaký trend, značí to, že uvažovaný model není schopen zcela zachytit vývoj úmrtnosti.

2.2 Renshaw-Habermanův model

Ve své práci [7] rozšiřují A. E. Renshaw a S. Haberman Lee-Carterův model o *kohortní efekt*. Kohortou se rozumí skupina osob, kterou spojuje stejná událost v daném časovém intervalu – v našem případě narození ve stejném kalendářním roce. Zavedení kohortního efektu bylo motivováno tím, že se mezi jednotlivými generacemi mění úmrtnost, důvodem může být různý přístup ke kouření, změna stravovacích návyků nebo obecně jiný životní styl.

Dle Renshawa a Habermana je možné věkově specifickou míru úmrtnosti modelovat jako

$$\log(m_x(t)) = a_1(x) + a_2(x) \cdot k_1(t) + a_3(x) \cdot c(t-x),$$

kde $a_1(x)$ označuje hlavní (statický) efekt věku a $a_2(x), a_3(x)$ tvoří spolu s $k_1(t)$ a $c(t-x)$ dynamické věkově specifické efekty roku a kohorty.

Všimněme si, že Lee-Carterův model je speciálním případem Renshaw-Habermanova modelu pro $a_3(x) = 0$ pro každý věk x. Naopak, pokud pro každý věk x položíme $a_2(x) = 0$, získáme *age-cohort* model, který nezahrnuje efekt kalendářního roku, ale pouze efekty věku a roku narození.

2.2.1 Odhad parametrů

Při odhadu parametrů se narozdíl od předchozího modelu, kdy jsme zdokumentovali v sekci 2.1.1 všechny nejpoužívanější metody odhadu, zaměřme pouze na odhad pomocí poissonovké odezvy.

Předpokládejme, že počet úmrtí ve věku x v roce t je náhodná veličina se střední hodnotou a rozptylem:

$$\begin{aligned} \mathsf{E}\ D_x(t) &= E_x(t) \cdot \exp\{a_1(x) + a_2(x) \cdot k_1(t) + a_3(x) \cdot c(t-x)\}, \\ \operatorname{var} D_x(t) &= \phi\, \mathsf{E}\ D_x(t) = \phi \frac{V(\mathsf{E}\ D_x(t))}{w_x(t)}, \end{aligned}$$

kde ϕ značí škálovací parametr, $V(\mathsf{E}\ D_x(t)) = \mathsf{E}\ D_x(t)$ je rozptylová funkce a pro váhy $w_x(t)$ platí, že $w_x(t) = 1$, pokud jsou příslušné vstupní údaje k dispozici (v opačném případě $w_x(t) = 0$).

Uvažujme zobecněný nelineární model pro počet úmrtí $D_x(t)$ – model s poissonovskou odezvou a logaritmickou linkovou funkcí, tj.

$$\log \mathsf{E}\ D_x(t) = \log E_x(t) + a_1(x) + a_2(x) \cdot k_1(t) + a_3(x) \cdot c(t-x).$$

Aby byly parametry modelu jednoznačně určeny, je třeba i v případě Renshaw-Habermanova modelu zavést omezující podmínky

$$\sum_{x=x_1}^{x_m} a_2(x) = 1, \quad \sum_{x=x_1}^{x_m} a_3(x) = 1$$

$$\text{a dále buď} \quad c(t_1 - x_m) = 0, \quad \text{nebo} \quad k_1(t_1) = 0.$$

Obdobně jako v případě Lee-Carterova modelu lze hodnotu parametru $a_1(x)$ odhadnout jako

$$\hat{a}_1(x) = \frac{1}{(t_n - t_1 + 1)} \sum_{t=t_1}^{t_n} \log(m_x(t)), \quad x = x_1, \ldots, x_m,$$

pro odhad ostatních parametrů navrhují autoři iterační postup založený na Newton-Raphsonově algoritmu, vzorce pro jednotlivé iterační kroky jsou uvedeny v [7].

2.2.2 Modelování a předpovídání časové složky

Na $k_1(t)$ a $c(t-x)$ je opět dle autorů [7] pohlíženo jako na realizaci časové řady řídící se modelem $ARIMA(p,d,q)$, opět se často jako vhodný model ukazuje náhodná procházka s driftem, tedy $ARIMA(0,1,0)$.

Projekci pro $t' > t_n$ je možné zapsat ve tvaru

$$\hat{m}_x(t') = \exp\{\hat{a}_1(x) + \hat{a}_2(x) \cdot \tilde{k}_1(t') + \hat{a}_3(x) \cdot \dot{c}(t'-x)\},$$

druhou možností je opět projektovat míru úmrtnosti na základě její poslední známé hodnoty $m_x(t_n)$ jako

$$\hat{m}_x(t') = m_x(t_n) \cdot \exp\{\hat{a}_2(x) \cdot (\tilde{k}_1(t') - \hat{k}_1(t_n)) + \hat{a}_3(x) \cdot (\dot{c}(t'-x) - \hat{c}(t_n - x))\}.$$

V obou možnostech výpočtu projektovaných hodnot označuje

$$\dot{c}(t'-x) = \begin{cases} \hat{c}(t'-x), & \text{je-li } 0 < t' - t_n \leq x - x_1, \\ \tilde{c}(t'-x), & \text{pro } t' - t_n > x - x_1, \end{cases}$$

kde $\hat{c}(t'-x)$ značí odhad získaný na základě pozorovaných hodnot pomocí metody uvedené v odstavci 2.2.1 a $\tilde{c}(t'-x)$ (stejně jako $\tilde{k}_1(t')$) zastupuje bodový odhad časové složky jako náhodného procesu.

2.2.3 Rezidua

Kvalitu odhadnutého modelu lze i v tomto případě vyhodnotit na základě grafického znázornění deviančních reziduí, která lze vypočítat dle vzorce

$$r_x(t) = sign(D_x(t) - \hat{D}_x(t)) \sqrt{\frac{dev(x,t)}{\hat{\phi}}},$$

kde

$$\begin{aligned} dev(x,t) &= 2w_x(t)\{D_x(t) \cdot \log(D_x(t)/\hat{D}_x(t)) - (D_x(t) - \hat{D}_x(t))\}, \\ \hat{\phi} &= \frac{\sum_{x,t} dev(x,t)}{\nu}, \\ \nu &= m(n-3) - 2(n-2). \end{aligned}$$

2.3 Cairns-Blake-Dowdův model

Cairns-Blake-Dowdův model (viz [8]), stejně jako Lee-Carterův model, patří do skupiny *age-period* modelů. Jedná se o model založený na představě, že ve vyšších věcích závisí $\log(q_x(t)/p_x(t))$ na x (pro pevně dané t) přibližně lineárně.

Cairns-Blake-Dowdův model se zpravidla uvádí ve tvaru, kdy odezvou je logistická transformace pravděpodobnosti úmrtí x-leté osoby (respektive logaritmus podílu pravděpodobností úmrtí a přežití x-letého jedince). Model CBD lze zapsat ve tvaru:

$$\operatorname{logit}(q_x(t)) = \log \frac{q_x(t)}{p_x(t)} = k_1(t) + x \cdot k_2(t),$$

kde $k_1(t), k_2(t)$ jsou časové řady. Intercept $k_1(t)$ obecně s časem klesá, což je dáno tím, že úmrtnost v jednotlivých věcích se časem snižuje.

2.3.1 Odhad parametrů

Narozdíl od Lee-Carterova a Renshaw-Habermanova modelu jsou parametry $k_1(t), k_2(t)$ určeny jednoznačně i bez dalších omezujících podmínek. Pro jejich odhad na základě pozorovaných pravděpodobností úmrtí uvažujme regresní model

$$\operatorname{logit}(q_x(t)) = \log \frac{q_x(t)}{p_x(t)} = k_1(t) + x \cdot k_2(t) + \varepsilon_{x,t},$$

kde $\varepsilon_{x,t}$ označuje náhodné odchylky – nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0, \sigma^2)$.

K odhadu parametrů modelu CBD je možné využít například běžnou metodu nejmenších čtverců. Pro každý pevně daný kalendářní rok t získáme hodnoty $\hat{k}_1(t), \hat{k}_2(t)$ minimalizací funkce

$$O^{LS}(k_1(t), k_2(t)) = \sum_{x=x_1}^{x_m} \left(\log \frac{q_x(t)}{p_x(t)} - k_1(t) - x \cdot k_2(t) \right)^2 .$$

2.3.2 Modelování a předpovídání časové složky

Na $\mathbf{k}(\mathbf{t}) = (k_1(t), k_2(t))^\top$ je dle [3] možné pohlížet jako na realizaci dvojrozměrné náhodné procházky s driftem ve tvaru

$$\begin{cases} k_1(t) = k_1(t-1) + d_1 + \xi_1(t), \\ k_2(t) = k_2(t-1) + d_2 + \xi_2(t), \end{cases}$$

kde symboly d_1, d_2 označují parametry driftu jednotlivých náhodných procházek a $\xi(t) = (\xi_1(t), \xi_2(t))^\top$ jsou nezávislé náhodné vektory s dvojrozměrným normálním rozdělením s nulovou střední hodnotou a kovarianční maticí Σ

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix} .$$

Parametry driftu d_1, d_2, marginální rozptyly σ_1^2, σ_2^2 a kovarianci $\sigma_{1,2}$ je možné odhadnout podle vzorců

$$\begin{aligned}
\hat{d}_i &= \frac{\hat{k}_i(t_n) - \hat{k}_i(t_1)}{t_n - t_1}, \qquad i = 1, 2, \\
\hat{\sigma}_i^2 &= \frac{1}{t_n - t_1} \sum_{t=t_2}^{t_n} \left(\hat{k}_i(t) - \hat{k}_i(t-1) - \hat{d}_i \right)^2, \qquad i = 1, 2, \\
\hat{\sigma}_{1,2} &= \frac{1}{t_n - t_1} \sum_{s=t_1}^{t_n-1} \sum_{t=t_1}^{t_n-1} \left(\hat{k}_1(s+1) - \hat{k}_1(s) - \hat{d}_1 \right) \left(\hat{k}_2(t+1) - \hat{k}_2(t) - \hat{d}_2 \right).
\end{aligned}$$

2.3.3 Rezidua

Pro zhodnocení správnosti uvažovaného modelu navrhuje [3] opět využít grafického znázornění standardizovaných Pearsonových reziduí, která lze v případě modelu CBD získat ze vztahu

$$r_x(t) = \frac{\log(q_x(t)/p_x(t)) - (\hat{k}_1(t) + x \cdot \hat{k}_2(t))}{\sqrt{\frac{1}{(x_m - x_1 - 1)(t_n - t_1 + 1)} \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} [\log(q_x(t)/p_x(t)) - (\hat{k}_1(t) + x \cdot \hat{k}_2(t))]^2}}.$$

3. Modelování vývoje úmrtnosti v korelovaných populacích

Stochastické modely úmrtnosti uvedené v kapitole 2 byly vytvořené pro modelování úmrtnosti v jedné populaci. V následujících podkapitolách uvedeme možnosti jejich rozšíření pro současné modelování úmrtnosti ve více populacích.

Úvahy o společném modelování úmrtnosti vycházejí podle [9] z představy, že pokud předpokládáme skupinu populací s podobnými socio-ekonomickými podmínkami a dalšími spojujícími faktory, která může být tvořena například populacemi mužů a žen stejné země, populacemi různých provincií nebo ras v jedné zemi nebo populacemi různých států z podobného regionu, neměly by se projektované úmrtnosti v populacích v čase z dlouhodobého hlediska vzdalovat. [10] formuluje tuto hypotézu matematicky pomocí podílu věkově specifických měr úmrtnosti v populacích i a j, tj. $m_x^{[i]}(t)/m_x^{[j]}(t)$, tento podíl by pro každý věk x neměl divergovat, když $t \to \infty$.

3.1 Rozšíření Lee-Carterova modelu

K popisu nejčastěji uvažovaných metod na zobecnění stochastických modelů úmrtnosti pro současné modelování ve více populacích využijme nejprve Lee-Carterův model. Problematikou rozšíření Lee-Carterova modelu se již v roce 2005 zabývali Li a Lee v [9], kde autoři popisují tzv. *model společných faktorů* a *rozšířený model společných faktorů*. Přehled metod uváděných v následujících odstavcích doplňují zobecnění pocházející zejména z [11].

Zatímco první tři níže uvažovaná zobecnění nesplňují nutnou podmínku pro konvergenci podílu věkově specifických měr úmrtnosti (viz dále vztah 3.4), v případě *modelu společných faktorů*, *rozšířeného modelu společných faktorů* a *modelu se společným* A_2 je možné vhodným projektováním časových řad reprezentujících efekt roku narození tuto konvergenci zaručit (možné způsoby projekce uvádí podkapitola 3.5.1).

3.1.1 Nezávislý model

Nejjednodušší metodou modelování úmrtnosti ve dvou populacích je využití samostatného Lee-Carterova modelu pro každou z populací, kde jednotlivé modely mezi sebou nemají žádný vztah. Matematicky můžeme takový model popsat jako

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + a_2^{[i]}(x) \cdot k_1^{[i]}(t) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2, \tag{3.1}$$

kde

$m_x^{[i]}(t)$ označuje specifickou míru úmrtnosti ve věku x v roce t v populaci i,

$a_1^{[i]}(x)$ je věkově specifický parametr průměrné úmrtnosti ve věku x v i-té populaci,

$k_1^{[i]}(t)$ zastupuje v čase se měnící index udávající úroveň úmrtnosti v populaci i v roce t,

$a_2^{[i]}(x)$ je parametr určující citlivost $\ln(m_x^{[i]}(t))$ na $k_1^{[i]}(t)$,

$\varepsilon_{x,t}^{[i]}$ jsou náhodné odchylky.

Parametr $a_1^{[i]}(x)$ v i-tém modelu je možné odhadnout jako aritmetický průměr $\ln(m_x^{[i]}(t))$ přes časový index a odhady parametrů $k_1^{[i]}(t)$ a $a_2^{[i]}(x)$ poté získat z historických dat pomocí singulárního rozkladu matice $\log(m_x^{[i]}(t)) - a_1^{[i]}(x)$ nebo dalších metod uvedených v sekci 2.1.1.

Obdobně jako v případě běžného Lee-Carterova modelu je v mnoha případech vhodné modelovat $k_1^{[i]}(t)$ pomocí náhodné procházky s driftem:

$$k_1^{[i]}(t) = k_1^{[i]}(t-1) + d^{[i]} + \xi_t^{[i]},$$

kde $d^{[i]}$ zastupuje drift pro jednotlivé populace a $\{\xi_t^{[i]}\}$ jsou posloupnosti nezávislých normálně rozdělených náhodných veličin, o kterých předpokládáme, že $\xi_t^{[1]}$ a $\xi_t^{[2]}$ jsou nezávislé. Pokud rezidua náhodné procházky nesplňují tyto předpoklady, je možné pro modelování řady $k_1^{[i]}(t)$ využít $ARIMA(p, d, q)$ proces s vyššími řády nebo jinou diferencí.

Tento přístup zanedbává jakoukoli závislost mezi mírami úmrtnosti ve zkoumaných populacích. Předpoklad nezávislosti je však těžko naplnitelný, v praxi budou úmrtnosti v blízkých populacích ovlivněny mnoha společnými faktory, může se jednat například o výskyt různých epidemií (případně pandemií), pokroky v léčbě závažných onemocnění, válečné konflikty v regionu nebo ekonomický vývoj oblasti. Uveďme proto dále modely, které závislost úmrtností mezi dvěma populacemi zohledňují.

3.1.2 Model se společným K_1

Model se společným K_1 předpokládá, že věkově specifické míry úmrtnosti ve dvou korelovaných populacích je možné modelovat pomocí Lee-Carterových modelů se společnými parametry $K_1(t) := k_1^{[1]}(t) = k_1^{[2]}(t), t = t_1, \ldots, t_n$, udávajícími hladinu úmrtnosti v roce t. Model pro i-tou populaci tedy lze zapsat ve tvaru

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + a_2^{[i]}(x) \cdot K_1(t) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2,$$

kde význam ostatních symbolů je shodný s modelem v předchozí sekci 3.1.1.

Parametry $a_1^{[i]}(x)$ je možné opět odhadnout jako aritmetický průměr $\ln(m_x^{[i]}(t))$ přes časový index v dané populaci. Pro získání odhadů parametrů $K_1(t)$ a $a_2^{[i]}(x)$ je poté autory navrženo použít singulární rozklad matice, která obsahuje hodnoty $\ln(m_x^{[i]}(t)) - \hat{a}_1^{[i]}(x)$ společně pro obě populace.

3.1.3 Kointegrovaný model

Možností, jak do *nezávislého modelu* z odstavce 3.1.1 vnést vztah mezi projektovanými mírami úmrtnosti ve dvou populacích, je pohlížet na náhodný vektor $\mathbf{k}_1(t) = (k_1^{[1]}(t), k_1^{[2]}(t))^\top$ jako na dvojrozměrnou náhodnou procházku, tj.

$$\mathbf{k}_1(t) = \mathbf{k}_1(t-1) + \mathbf{d} + \xi_t, \tag{3.2}$$

kde $\mathbf{d} = (d^{[1]}, d^{[2]})^\top$ je vektor parametrů driftu a $\xi_t = (\xi_t^{[1]}, \xi_t^{[2]})^\top$ je posloupnost nekorelovaných náhodných vektorů s dvojrozměrným normálním rozdělením s nulovou střední hodnotou a kovarianční maticí Σ. Právě prostřednictvím matice Σ je v modelu zohledněna závislost mezi populacemi.

V praxi můžeme často pozorovat kointegraci mezi procesy, tedy, jak uvádějí Li a Hardy ve své práci [11], situaci, kdy oba procesy $\{k_1^{[1]}(t)\}$ a $\{k_1^{[2]}(t)\}$ jsou nestacionární a zároveň existuje konstanta β taková, že časová řada $\{k_1^{[1]}(t) - \beta k_1^{[2]}(t)\}$ je stacionární (ve smyslu slabé stacionarity). Proto navrhují pro identifikaci existence kointegrace mezi dvěma procesy a pro odhad konstanty β následující algoritmus:

1. Použít rozšířený Dickey-Fullerův test pro ověření nestacionarity každého z procesů $\{k_1^{[1]}(t)\}$, $\{k_1^{[2]}(t)\}$ založený na vztahu

   $$\Delta k_1^{[i]}(t) = \sum_{j=1}^{p} \Delta k_1^{[i]}(t-j) + \gamma k_1^{[i]}(t-1) + \xi(t),$$

   kde $\Delta k_1^{[i]}(t) = k_1^{[i]}(t) - k_1^{[i]}(t-1)$, p značí počet zpoždění a $\xi(t)$ náhodnou odchylku. Tento test testuje nulovou hypotézou, že $\gamma = 0$, tedy že proces je nestacionární, proti alternativě, že $\gamma < 0$. Testovou statistikou je tzv. t-poměr: $t_\gamma = \hat{\gamma}/sd(\hat{\gamma})$, ve kterém $\hat{\gamma}$ značí odhad parametru γ metodou nejmenších čtverců v uvedeném vztahu a $sd(\hat{\gamma})$ jeho směrodatnou odchylku. Kritické hodnoty pro tento test jsou tabelovány. Více o testu uvádí [12]. Pokud nezamítáme H_0 pro oba procesy, je možné pokračovat dalším krokem algoritmu. V opačném případě nelze považovat procesy za kointegrované.

2. Odhadnout konstantu β v lineárním regresním modelu

   $$k_1^{[1]}(t) = \alpha + \beta \cdot k_1^{[2]}(t) + u_t,$$

   kde α je konstanta.

3. Testovat stacionaritu reziduální složky $\{u_t\}$ z kroku 2 pomocí rozšířeného Dickey-Fullerova testu. Pokud tento test zamítne hypotézu nestacionarity, je možné považovat procesy $\{k_1^{[1]}(t)\}$, $\{k_1^{[2]}(t)\}$ za kointegrované.

V případě, že shledáme, že procesy $\{k_1^{[1]}(t)\}$ a $\{k_1^{[2]}(t)\}$ jsou kointegrované, je možné oba procesy modelovat pouze pomocí jedné (jednorozměrné) náhodné procházky

$$\begin{aligned} k_1^{[1]}(t) &= k_1^{[1]}(t-1) + d + \xi_t^{[1]}, \\ k_1^{[2]}(t) &= a + b \cdot k_1^{[1]}(t) + u_t, \end{aligned} \tag{3.3}$$

kde a, b a d jsou konstanty a $\xi_t^{[1]}$ a u_t jsou náhodné odchylky s normálním rozdělením s nulovou střední hodnotou.

3.1.4 Model společných faktorů

Postačující podmínkou pro to, aby podíl projektovaných měr úmrtnosti dvou populací nedivergoval, je dle [10]

$$a_2^{[1]}(x) = a_2^{[2]}(x) \qquad \text{pro všechna } x \quad \text{a} \tag{3.4}$$

$$k_1^{[1]}(t) - k_1^{[2]}(t) \quad \text{je „mean-reverting“}, \tag{3.5}$$

kde procesem s vlastností *mean-reverting* rozumíme proces, který má tendencí se navracet k dlouhodobé střední úrovni. Proces se spojitým časem $\{Y_t\}$ s vlastností *mean-reversion* podle [13] vyhovuje stochastické diferenciální rovnici

$$dY_t = \theta(K - Y_t)dt + \sigma dW_t,$$

kde $\theta > 0$, K a σ jsou parametry a W_t označuje Wienerův proces. Pro účely modelování $k_1^{[1]}(t) - k_1^{[2]}(t)$ je však vhodné uvažovat analogii takového procesu pro diskrétní čas, kterou je dle [13] autoregresní proces prvního řádu $AR(1)$

$$X_t = K(1 - e^{-\theta}) + e^{-\theta} X_{t-1} + \sigma\sqrt{(1 - e^{-2\theta})/2\theta} \cdot \varepsilon_t,$$

kde $0 < e^{-\theta} < 1$ a ε_t jsou nezávislé veličiny s rozdělením $N(0,1)$. Jednu z možností, jak lze projektovat efekt kalendářního roku ve dvou populacích takovým způsobem, aby byla splněna uvedená podmínka, popisuje podkapitola 3.5.1.

Pokud bychom modelovali efekty kalendářních roků ($k_1^{[1]}(t)$ a $k_1^{[2]}(t)$) pomocí náhodných procházek s driftem, bylo by podle [9] nutné pro splnění podmínky, aby podíl očekávaných měr úmrtnosti v populacích nedivergoval, požadovat rovnost $a_2^{[1]}(x)$ a $a_2^{[2]}(x)$ a dále shodný parametr driftu náhodných procházek. V takovém případě by byl uvedený podíl v čase konstantní pro každý věk x.
Jelikož je dle autorů v praxi nepravděpodobné, aby dva různé procesy $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ měly stejné driftové parametry, navrhuje [9] použít model, ve kterém

$$\begin{aligned} a_2^{[1]}(x) = a_2^{[2]}(x) &=: A_2(x) \quad \text{pro všechna } x, \\ k_1^{[1]}(t) = k_1^{[2]}(t) &=: K_1(t) \quad \text{pro všechna } t. \end{aligned}$$

Tento model lze tedy zapsat ve tvaru

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + A_2(x) \cdot K_1(t) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2,$$

kde $\varepsilon_{x,t}^{[i]}$ jsou náhodné odchylky. Součin parametrů $A_2(x)$ a $K_1(t)$ v takovém případě zachycuje společné změny v úmrtnosti obou populací.

3.1.5 Rozšířený model společných faktorů

Li a Lee v [9] dále zobecňují *model společných faktorů* přidáním populačně-specifických faktorů $a^{[i]}(x)$ a $k^{[i]}(t)$ do modelu, ten lze poté zapsat jako

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + A_2(x) \cdot K_1(t) + a^{[i]}(x) \cdot k^{[i]}(t) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2.$$

Pro odhad parametrů $a_1^{[i]}(x)$, $A_2(x)$ a $K_1(t)$ je možné, stejně jako v případě *modelu společných faktorů*, využít například metodou nejmenších čtverců, tedy minimalizaci

$$\min \sum_{t=t_1}^{t_n} [\log(m_x^{[i]}(t)) - a_1^{[i]}(x) - A_2(x) \cdot K_1(t)]^2,$$

odkud lze po použití podmínky pro jednoznačnost parametrů Lee-Carterova modelu ($\sum_{t=t_1}^{t_n} K_1(t) = 0$) získat odhad $a_1^{[i]}(x)$:

$$\hat{a}_1^{[i]}(x) = \frac{\sum_{t=t_1}^{t_n} \log(m_x^{[i]}(t))}{t_n - t_1 + 1}.$$

Odhadnout parametry $A_2(x)$ a $K_1(t)$ je poté možné použitím singulárního rozkladu matice, která obsahuje hodnoty $\sum_{i=1}^{2} [\log(m_x^{[i]}(t)) - \hat{a}_1^{[i]}(x)]$.

Odhady parametrů $a^{[i]}(x)$ a $k^{[i]}(t)$ rozšířeného modelu se následně získají singulárním rozkladem matice reziduí i-té populace (tj. reziduí, která vzejdou z *modelu společných faktorů* a příslušejí i-té populaci).

Narozdíl od ostatních verzí Lee-Carterova modelu doporučují autoři pro projekci specifické míry úmrtnosti modelovat řady $k^{[i]}(t)$ jako $AR(1)$ proces (namísto obvykle užívané náhodné procházky s driftem)

$$k^{[i]}(t) = c_0^{[i]} + c_1^{[i]} \cdot k^{[i]}(t-1) + \xi_t^{[i]},$$

kde $c_0^{[i]}, c_1^{[i]}$ jsou konstanty a $\xi_t^{[i]}$ označuje nezávislé stejně rozdělené náhodné chyby s nulovou střední hodnotou. Od konstant $c_1^{[i]}$ dále požadujeme, aby $|c_1^{[i]}| < 1$ pro obě populace.

3.1.6 Model se společným A_2

Poslední z možných rozšíření Lee-Carterova modelu pro modelování úmrtnosti více populací, které budeme v této práci uvažovat, uvádí [10]. Jedná se o model, ve kterém pro obě populace uvažujeme pouze shodné věkově specifické parametry $a_2^{[1]}(x) = a_2^{[2]}(x) =: A_2(x)$.

Model pro i-tou populaci tedy je tedy možné zapsat jako

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + A_2(x) \cdot k_1^{[i]}(t) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2.$$

Stejně jako u *modelu společných faktorů* a *rozšířeného modelu společných faktorů* lze i v tomto modelu zajistit, aby se projektované věkově specifické míry úmrtnosti ve dvou populacích nevzdalovaly. V tomto případě je možné tohoto dosáhnout, pokud pro projekci časových řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ využijeme proces RWAR, který je blíže popsán v podkapitole 3.5.

3.2 Renshaw-Habermanův model pro více populací

Stejně jako v případě Lee-Carterova modelu je možné uvažovat několik možných způsobů společného modelování úmrtnosti ve dvou populacích pomocí Renshaw-Habermanova modelu. Uvažujme obecně model

$$\log(m_x^{[i]}(t)) = a_1^{[i]}(x) + a_2^{[i]}(x) \cdot k_1^{[i]}(t) + a_3^{[i]}(x) \cdot c^{[i]}(t-x) + \varepsilon_{x,t}^{[i]}, \qquad i = 1, 2.$$

Postačující podmínkou pro konvergenci podílu projektovaných měr úmrtnosti ve dvou populacích je dle [10]

$$\begin{aligned} a_2^{[1]}(x) &= a_2^{[2]}(x), \\ a_3^{[1]}(x) &= a_3^{[2]}(x), \\ k_1^{[1]}(t) - k_1^{[2]}(t) & \quad \text{je „mean-reverting“ a} \\ c^{[1]}(t-x) - c^{[2]}(t-x) & \quad \text{je „mean-reverting“} \end{aligned}$$

a obdobně jako v případě Lee-Carterova modelu ji splňuje například model, ve kterém uvažujeme rovnost mezi parametry zmíněnými v každém řádku této podmínky – *model společných faktorů.*

V případě Renshaw-Habermanova modelu je možné mimo nezávislého modelu uvažovat i modely se spojenými některými z časových řad:

model se společným K_1:	$k_1^{[1]}(t) = k_1^{[2]}(t) =: K_1(t)$,
model společné kohorty :	$c^{[1]}(t-x) = c^{[2]}(t-x) =: C(t-x)$,
model společné kohorty a společného K_1:	$k_1^{[1]}(t) = k_1^{[2]}(t) =: K_1(t)$,
	$c^{[1]}(t-x) = c^{[2]}(t-x) =: C(t-x)$,

pokud zároveň požadujeme rovnost parametrů $a_2^{[i]}(x)$ pro $i = 1, 2$, respektive $a_3^{[i]}(x)$ pro $i = 1, 2$, tj.

$$\begin{aligned} a_2^{[1]}(x) = a_2^{[2]}(x) =: A_2(x) & \quad \text{pro všechna } x, \\ a_3^{[1]}(x) = a_3^{[2]}(x) =: A_3(x) & \quad \text{pro všechna } x, \end{aligned}$$

lze projektovat věkově specifickou míru úmrtnosti v populacích tak, aby její podíl pro každý věk x nedivergoval, a to pomocí modelů pro efekt kalendářního roku a kohortní efekt, které uvádí dále podkapitola 3.5.

3.3 Zobecnění Cairns-Blake-Dowdova modelu

V případě Cairns-Blake-Dowdova modelu lze zobecnění pro současné modelování úmrtnosti ve dvou populacích zapsat ve tvaru:

$$\operatorname{logit}(q_x^{[i]}(t)) = \log \frac{q_x^{[i]}(t)}{p_x^{[i]}(t)} = k_1^{[i]}(t) + x \cdot k_2^{[i]}(t), \qquad i = 1, 2.$$

Postačujícími podmínkami pro konvergenci podílu projektovaných měr úmrtnosti ve dvou populacích jsou v případě tohoto modelu podmínky:

$$\begin{aligned} k_1^{[1]}(t) - k_1^{[2]}(t) &\quad \text{je „mean-reverting“ a} \\ k_2^{[1]}(t) - k_2^{[2]}(t) &\quad \text{je „mean-reverting“.} \end{aligned}$$

Tyto podmínky jsou rovněž splněny například v modelu, ve kterém platí rovnost mezi parametry v řádcích podmínek, tj.

$$\begin{aligned} k_1^{[1]}(t) &= k_1^{[2]}(t) =: K_1(t) \\ k_2^{[1]}(t) &= k_2^{[2]}(t) =: K_2(t). \end{aligned}$$

Dále je pro současné modelování úmrtnosti ve více populacích i v případě Cairns-Blake-Dowdova modelu možné uvažovat nezávislý model, případně model se společným K_1, nebo model se společným K_2. V případě těchto modelů je pro splnění podmínky konvergence podílu měr úmrtnosti dle [10] rovněž vyžadována projekce časových řad pomocí modelů uvedených v podkapitole 3.5.

3.4 Odhad metodou maximální věrohodnosti

Odhadnout parametry modelů uvedených v předchozích podkapitolách (respektive jejich zobecnění pro p zkoumaných populací), tedy modelů

$$\begin{aligned} LC : \log(m_x^{[i]}(t)) &= a_1^{[i]}(x) + a_2^{[i]}(x) \cdot k_1^{[i]}(t), i = 1, 2, \ldots, p, \\ RH : \log(m_x^{[i]}(t)) &= a_1^{[i]}(x) + a_2^{[i]}(x) \cdot k_1^{[i]}(t) + \\ &+ a_3^{[i]}(x) \cdot c^{[i]}(t - x), \quad i = 1, 2, \ldots, p, \\ CBD : \operatorname{logit}(q_x^{[i]}(t)) &= \log \frac{q_x^{[i]}(t)}{p_x^{[i]}(t)} = k_1^{[i]}(t) + x \cdot k_2^{[i]}(t), i = 1, 2, \ldots, p, \end{aligned} \tag{3.6}$$

je možné za předpokladu poissonovsky rozdělených počtů úmrtí $D_x^{[i]}(t)$ v i-té populaci ve věku x a v roce t, tj.

$$D_x(t) \sim Po\left(E_x(t) \cdot m_x^{[i]}(t)\right),$$

s využitím metody maximální věrohodnosti. Funkci, kterou chceme maximalizovat, je možné dle [10] obecně zapsat ve tvaru

$$\begin{aligned} \ell = \sum_{i=1}^{p} \sum_{x=x_1}^{x_m} \sum_{t=t_1}^{t_n} w_x^{[i]}(t) \{ D_x^{[i]}(t) \cdot \log(E_x^{[i]}(t) \cdot m_x^{[i]}(t)) - \\ - E_x^{[i]}(t) \cdot m_x^{[i]}(t) - \log(D_x^{[i]}(t)!) \}, \end{aligned} \tag{3.7}$$

kde $E_x^{[i]}(t)$ je příslušná expozice vůči riziku úmrtí ve věku x v roce t v i-té populaci. Váhy $w_x^{[i]}(t)$ je možné využít k vyloučení nežádoucích kohort z datového souboru (např. první čtyři a poslední čtyři kohorty).

V obecném zápise modelů 3.6 je možné nahradit (respektive sjednotit) některé parametry s ohledem na výše zmíněné možnosti zobecnění stochastických modelů

úmrtnosti pro modelování více populací. Následně je možné získat odhady parametrů položením parciálních derivací funkce ℓ podle příslušných parametrů rovné nule a řešením získaného vztahu například s využitím Newton-Raphsonovy iterační metody.

Na závěr každé iterace je třeba pro Lee-Carterův a Renshaw-Habermanův model zohlednit podmínky pro jednoznačnost řešení, obdobně jako v případě modelů pro jednu populaci se jedná o podmínky:

$$\sum_{x=x_1}^{x_m} a_2^{[i]}(x) = 1 \quad \text{a} \quad \sum_{t=t_1}^{t_n} k_1^{[i]}(t) = 0, \quad i = 1, 2, \ldots, p$$

v případě Lee-Carterova modelu a

$$\sum_{x=x_1}^{x_m} a_2^{[i]}(x) = 1, \quad \sum_{x=x_1}^{x_m} a_3^{[i]}(x) = 1$$

$$\text{a dále buď} \quad c^{[i]}(t_1 - x_m) = 0 \quad \text{nebo} \quad k_1^{[i]}(t_1) = 0, \quad i = 1, 2, \ldots, p$$

v případě Renshaw-Habermanova modelu.

3.5 Projekce časových řad

Pro projekci časových řad reprezentujících efekt kalendářního roku a efekt roku narození (kohortní efekt) ve dvou populacích navrhuje [10] využít níže uvedené modely časových řad.

Modely se liší v závislosti na tom, zda se jedná o modelování časové řady efektu spojeného pro obě populace nebo zbývajících efektů, které jsou pro jednotlivé populace rozdílné.

Jelikož se budeme zaměřovat na projekci úmrtnosti ve stejném věkovém rozsahu, jaký byl uvažován pro odhad parametrů, není třeba se zabývat extrapolací parametrů udávajících závislost míry úmrtnosti na věku, tj. parametrů $a_1^{[i]}(x)$, $a_2^{[i]}(x)$ a $a_3^{[i]}(x)$.

3.5.1 Projekce efektu kalendářního roku

Pro modelování efektů kalendářního roku samostatných pro jednotlivé populace navrhují autoři použít *náhodnou procházku s autoregresním procesem prvního řádu*, označme $RWAR$. Při tomto přístupu je efekt kalendářního roku první populace $k_i^{[1]}(t)$ modelován jako náhodná procházka s driftem, zatímco rozdíl efektů mezi dvěma populacemi $k_i^{[1]}(t) - k_i^{[2]}(t)$ je modelován pomocí autoregresního procesu prvního řádu. Dolní index i je v případě Lee-Carterova a Renshaw-Habermanova modelu roven 1, v případě Cairns-Blake-Dowdova modelu nabývá hodnot 1 a 2.

Model RWAR lze tedy zapsat ve tvaru

$$\begin{aligned} k_i^{[1]}(t) &= \mu_i^{[1]} + k_i^{[1]}(t-1) + Z_i^{[1]}(t), \\ k_i^{[1]}(t) - k_i^{[2]}(t) &= \mu_i^{[2]} + \phi_i^{[2]} \cdot (k_i^{[1]}(t-1) - k_i^{[2]}(t-1)) + Z_i^{[2]}(t), \end{aligned}$$

kde $\mu_i^{[1]}$, $\mu_i^{[2]}$ a $\phi_i^{[2]}$ jsou parametry modelu a $Z_i^{[1]}(t)$ a $Z_i^{[2]}(t)$ jsou náhodné odchylky. O vektoru $(Z_i^{[1]}(t), Z_i^{[2]}(t))$ předpokládáme, že má dvojrozměrné normální rozdělení s nulovou střední hodnotou a konstantní kovarianční maticí a že pro $t \neq u$ platí, že vektory $(Z_i^{[1]}(t), Z_i^{[2]}(t))$ a $(Z_i^{[1]}(u), Z_i^{[2]}(u))$ jsou nezávislé.
Od parametru $\phi_i^{[2]}$ vyžadujeme, aby splňoval $0 < \phi_i^{[2]} < 1$, čímž zaručíme, že rozdíl procesů $k_i^{[1]}(t) - k_i^{[2]}(t)$ je *mean-reverting*.

Pokud je pro modelování samostatných efektů kalendářního roku využit model $RWAR$, je možné pro modelování spojených řad $K_i(t)$ použít náhodnou procházku s driftem, tj.

$$K_i(t) = \mu_i + K_i(t-1) + Z_i(t),$$

kde μ_i je parametr driftu a $Z_i(t)$ jsou nezávislé normálně rozdělené náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem.

3.5.2 Projekce kohortního efektu

K modelování efektů roku narození (kohortních efektů) v případě, kdy jsou příslušné časové řady samostatné pro jednotlivé populace je navrhováno použít *gravitační model*, ve kterém pro efekty roku narození $y = t - x$ v jednotlivých populacích, tj. $c^{[1]}(y)$ a $c^{[2]}(y)$, platí

$$\begin{aligned} c^{[1]}(y) - c^{[1]}(y-1) &= \mu^{[1]} + \phi^{[1]} \cdot (c^{[1]}(y-1) - c^{[1]}(y-2)) + Z^{[1]}(y), \\ c^{[2]}(y) - c^{[2]}(y-1)) &= \mu^{[2]} + \alpha^{[2]} \cdot (c^{[1]}(y-1) - c^{[2]}(y-1)) + \\ &+ \phi_i^{[2]} \cdot (c^{[2]}(y-1) - c^{[2]}(y-2)) + Z^{[2]}(y), \end{aligned}$$

kde $\mu^{[1]}$, $\phi^{[1]}$, $\mu^{[2]}$, $\alpha^{[2]}$ a $\phi^{[2]}$ jsou parametry modelu a $Z^{[1]}(y)$ a $Z^{[2]}(y)$ jsou náhodné odchylky s dvojrozměrným normálním rozdělením s nulovou střední hodnotou a konstantní kovarianční maticí. Od parametru $\alpha^{[2]}$ dále požadujeme, aby $|\alpha^{[2]}| < 1$, z důvodu, aby se rozdíl v projekcích kohortních efektů nezvyšoval.

Pro modelování spojených kohortních efektů $C(y)$ je navrženo použít proces $ARIMA(p, d, q)$, tj.

$$\begin{aligned} \Delta^d C(y) &= \mu + \phi_1 \Delta^d C(y-1) + \phi_2 \Delta^d C(y-2) + \cdots + \phi_p \Delta^d C(y-p) + \\ &+ Z(y) + \psi_1 Z(y-1) + \psi_2 Z(y-2) + \cdots + \psi_q Z(y-q), \end{aligned}$$

s diferencí d a řády p a q takovými, aby hodnota Bayesova informačního kritéria tohoto procesu byla co nejlepší. V rovnici procesu označují μ, $\phi_1, \ldots, \phi_p$ a $\psi_1, \ldots, \psi_q$ parametry modelu a $Z(y)$ nezávislé normálně rozdělené náhodné veličiny s nulovou střední hodnotou a konstantním rozptylem.

4. Aplikace modelů na reálná data

Cílem této části práce je aplikovat stochastické modely úmrtnosti pro více populací z kapitoly 3 na reálná data a ilustrovat rozdílnost predikcí plynoucích z těchto modelů v porovnání s modely pro samostatné populace. Pro výpočty byly využity údaje o úmrtnosti v České a Slovenské republice v letech 1970 – 2014. Použitá data jsou v obou případech převzata z *Human Mortality Database* [14], která shromažďuje údaje o počtech narození, velikostech populace a o úmrtnosti ve vybraných státech světa (aktuálně se jedná o 38 států).

Z *Human mortality database* byla využita souhrnná data pro obě pohlaví, která jsou váženým průměrem zohledňujícím zastoupení jednotlivých pohlaví v daném věku a kalendářním roce v populaci. Zveřejňované údaje jsou dále upraveny tak, aby byla například doplněna chybějící data, bližší informace o metodice úprav dat je možné nalézt v [15].

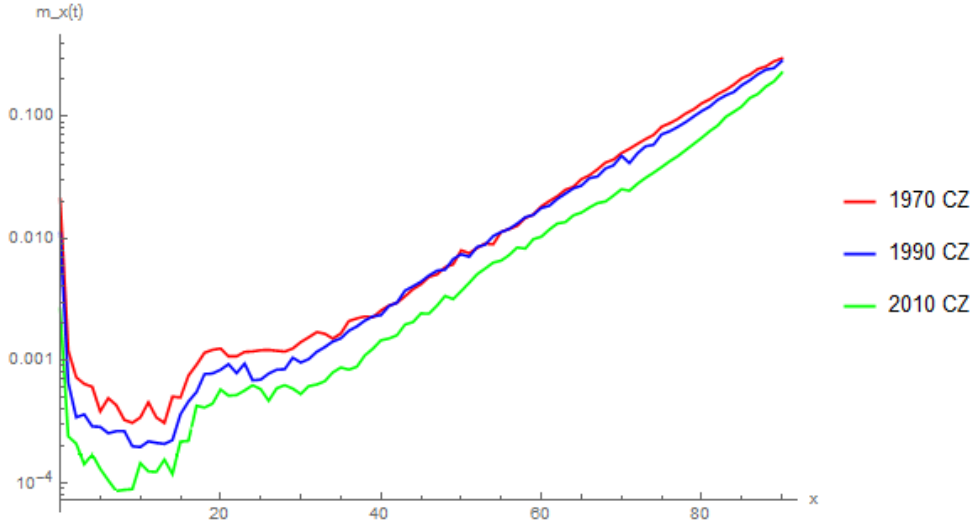
Pro výpočty byl využit software Wolfram Mathematica, použitý zdrojový kód je uložen na přiloženém CD. Projekce časových řad a testování jejich vlastností (resp. vlastností reziduálních složek) bylo provedeno v systému EViews, soubor je rovněž uložen na CD. Uvažované testy byly prováděny na hladině spolehlivosti 5 %.

4.1 Úmrtnost české populace

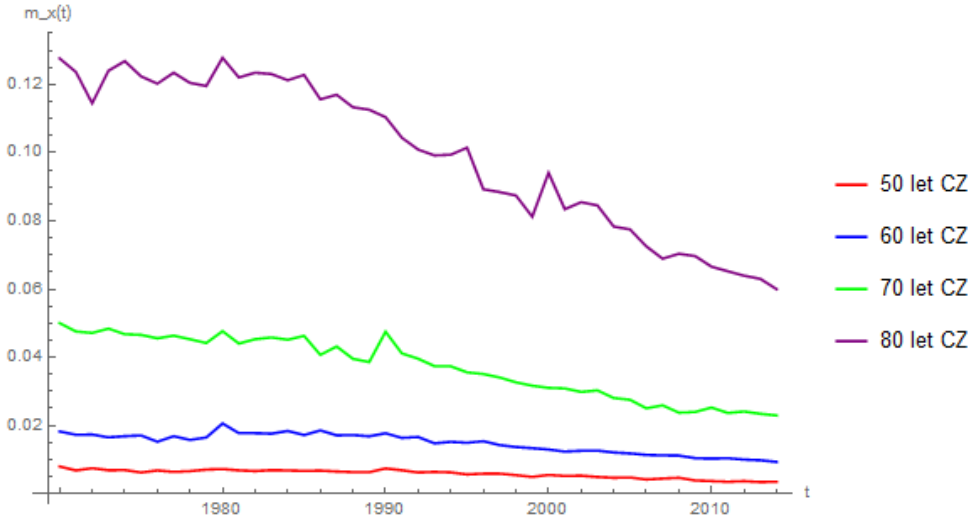
Věnujme se nejprve úmrtnosti české populace. Obrázek 4.1 znázorňuje v logaritmickém měřítku věkově specifickou míru úmrtnosti v České republice v letech 1970, 1990 a 2010. V obrázku je zřejmá poměrně vysoká úmrtnost v prvním roce života, která souvisí s rizikem úmrtí při porodu a kojeneckých nemocech. Přibližně do patnáctého roku je potom míra úmrtnosti velmi nízká. Od puberty však přibližně do 25 let zaznamenáváme značný nárůst úmrtnosti, který souvisí s nehodovostí, neopatrným užíváním drog a alkoholu a chronickými chorobami postihujícími osoby v tomto věku, tento skokovitý nárůst úmrtnosti obvykle nazýváme "accident hump". Přibližně od třiceti let poté úmrtnost exponenciálně narůstá bez významnějších výkyvů. Z obrázku vidíme, že v nižším věku se s kalendářním rokem významně snižuje úmrtnost a zejména po roce 1990 dochází ke snižování úmrtnosti ve všech věcích. Detailnější zachycení vývoje specifické míry úmrtnosti v České republice ve věcích 50, 60, 70 a 80 let v průběhu období 1970 – 2014 poskytuje obrázek 4.2.

Jak bylo uvedeno dříve, použití Cairns-Blake-Dowdova modelu je vhodné pro modelování úmrtnosti ve vyšších věcích, proto se s odhadem stochastických modelů omezme na úmrtnost ve věku 45 – 90 let v letech 1970 – 2014.

K modelování úmrtnosti využijme stochastických modelů uvedených v kapitole 2, tedy Lee-Carterova a Renshaw-Habermanova modelu pro věkově spe-



Obrázek 4.1: Specifická míra úmrtnosti v letech 1970, 1990 a 2010, ČR



Obrázek 4.2: Specifická míra úmrtnosti ve věku 50, 60, 70 a 80 let, ČR

cifickou míru úmrtnosti $m_x(t)$, $x = 45, \ldots, 90$, $t = 1970, \ldots, 2014$ a Cairns--Blake-Dowdova modelu pro logistickou transformaci pravděpodobnosti úmrtí $q_x(t), x = 45, ..., 90$, $t = 1970, ..., 2014$.

Obrázek 4.3 znázorňuje postupně standardizovaná Pearsonova rezidua Lee--Carterova modelu, devianční rezidua Renshaw-Habermanova a Pearsonova rezidua Cairns-Blake-Dowdova modelu pro úmrtnost české populace. Vyobrazena jsou rezidua modelu pro věk 60 a 80 let. Jak vidíme z obrázku, pro modelování úmrtnosti je nejméně vhodný Cairns-Blake-Dowdův model, ve kterém je zřejmý trend reziduí po roce 1980 (rostoucí v případě věku 60 let a klesající pro 80 let). Naproti tomu rezidua Lee-Carterova a Renshaw- -Habermanova modelu je možné dle tohoto obrázku považovat za náhodně rozmístěná.

Pro rozhodnutí, který model je pro modelování úmrtnosti v České republice nejvhodnější, využijme logaritmické věrohodnostní funkce – za předpokladu, že počty úmrtí $D_x(t)$ v každém věku $x = 45, \ldots, 90$ a roce $t = 1970, \ldots, 2014$ mají Poissonovo rozdělení se střední hodnotou $E_x(t) \cdot m_x(t)$. Abychom penalizovali modely s větším počtem parametrů, rozhodněme o nejvhodnějším modelu na základě hodnoty Bayesova informačního kritéria (BIC)

$$BIC_m = \ell(\hat{\phi}_m) - \frac{1}{2}\nu_m \cdot \log N, \tag{4.1}$$

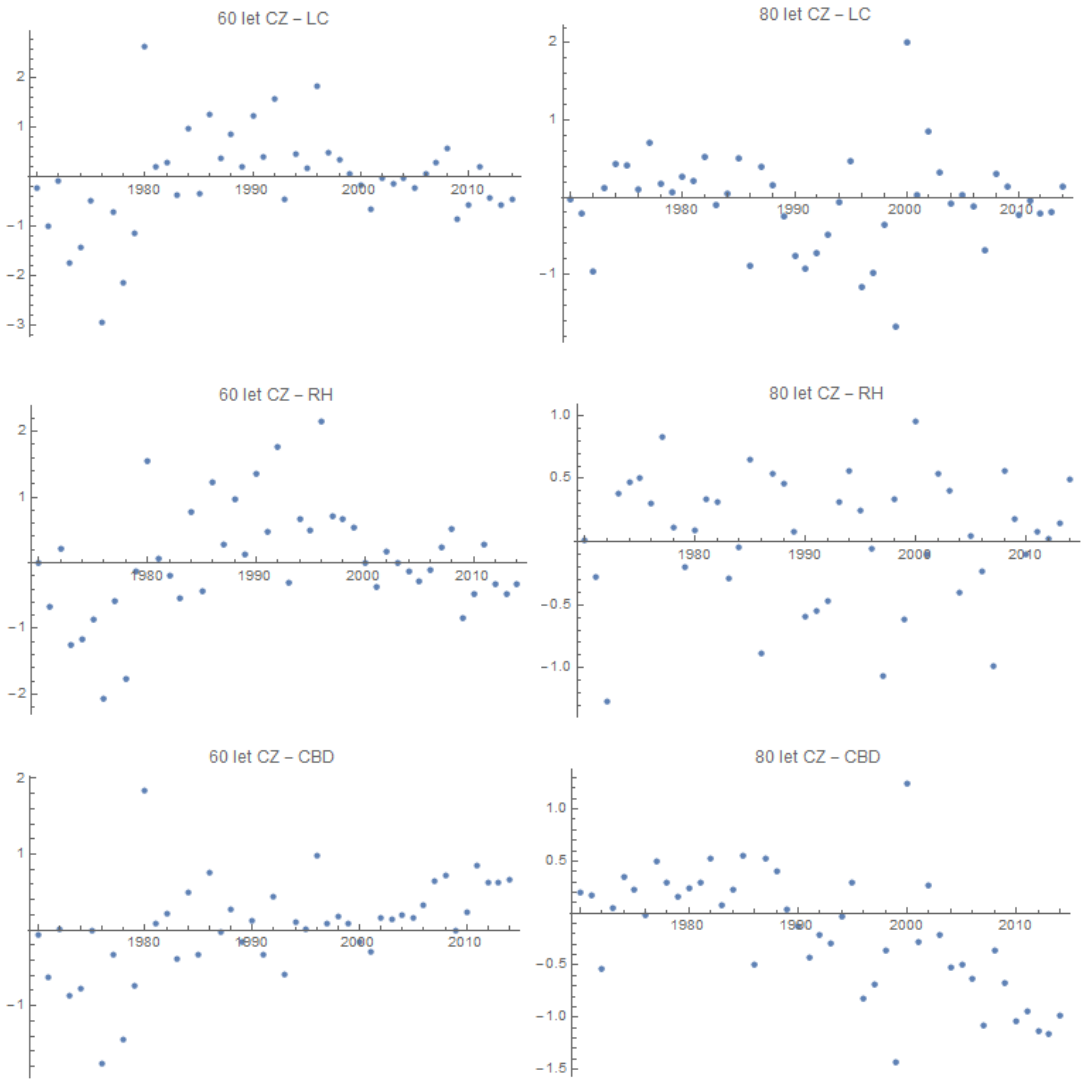
kde ϕ_m je vektor parametrů modelu m, $\hat{\phi}_m$ je jejich maximálně věrohodný odhad, $\ell(\hat{\phi}_m)$ je hodnota logaritmické věrohodnostní funkce v $\hat{\phi}_m$, ν_m označuje počet odhadovaných parametrů snížený o počet podmínek a N značí počet pozorování, na jejichž základě je model odhadnut. Při takto definovaném BIC preferujeme modely s vyšší hodnotou informačního kritéria.

Tabulka číslo 4.1 uvádí pro každý z uvažovaných modelů maximum logaritmické věrohodnostní funkce, počet odhadovaných parametrů a hodnotu BIC. Odhady parametrů modelů vycházejí z údajů o úmrtnosti ve 46 věcích a 45 kalendářních rocích, hodnota N je tedy ve všech případech rovna 2070. Vidíme, že nejvyššího BIC nabývá i přes penalizaci za větší počet parametrů Renshaw--Habermanův model, který předstihl Lee-Carterův model. Obdobně jako v grafickém znázornění reziduí (obrázek 4.3) se Cairns-Blake-Dowdův model ukazuje pro modelování úmrtnosti v České republice jako nejméně vhodný.

Model	**Maximální věrohodnost**	**Počet parametrů**	**BIC**	**Pořadí**
Lee-Carter	−12 921,1	135	−13 436,5	2
Renshaw-Haberman	−11 638,9	270	−12 681,1	1
Cairns-Blake-Dowd	−18 322,4	90	−18 666,0	3

Tabulka 4.1: Odhad parametrů metodou maximální věrohodnosti – ČR

Jelikož Lee-Carterův model je speciálním případem Renshaw-Habermanova modelu, ve kterém $a_3(x) = 0$ pro každý věk x, je možné dále testovat hypotézu, že



Obrázek 4.3: Rezidua modelů úmrtnosti pro věk 60 a 80 let, ČR

Lee-Carterův model popisuje úmrtnost české populace stejně dobře jako Renshaw--Habermanův model, proti alternativě, že Renshaw-Habermanův model je pro modelování úmrtnosti české populace signifikantně lepší. K tomuto účelu je možné využít *likelihood-ratio* test, jehož testová statistika je založena na rozdílu maxim logaritmických věrohodnostních funkcí

$$2 \cdot (\ell(\hat{\phi}_{R-H}) - \ell(\hat{\phi}_{L-C}))$$

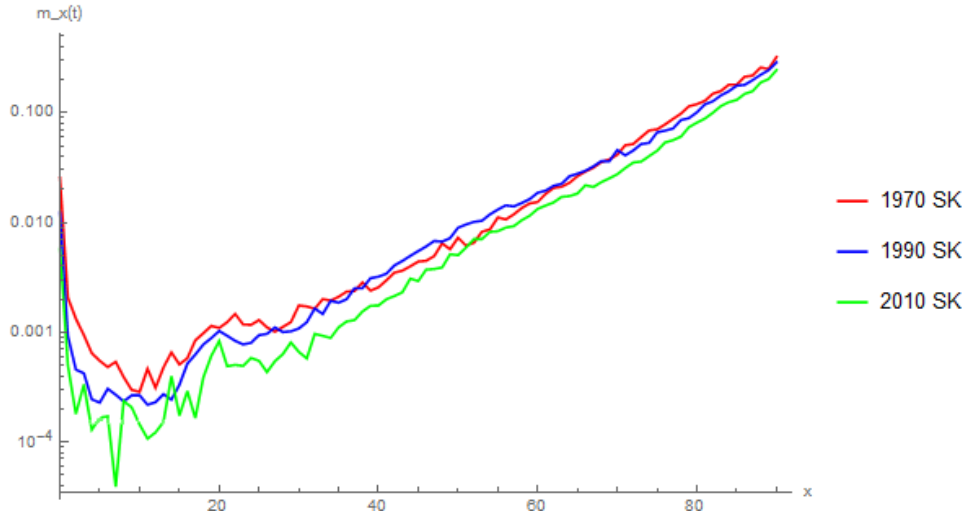
a za platnosti nulové hypotézy má asymptoticky rozdělení χ^2 s počtem stupňů volnosti rovným rozdílu v počtu parametrů modelů. Jelikož pro využitá data o úmrtnosti v české populaci nabývá testová statistika hodnoty 2 564,4 a je tedy výrazně vyšší než kritická hodnota testu $\chi^2_{135}(0{,}95) = 163{,}12$, zamítáme (na hladině spolehlivosti 5 %) nulovou hypotézu a můžeme konstatovat, že z uvažovaných modelů pro úmrtnost ve věku $45 - 90$ let v české populaci je nejvhodnějším modelem Renshaw-Habermanův.

4.2 Úmrtnost slovenské populace

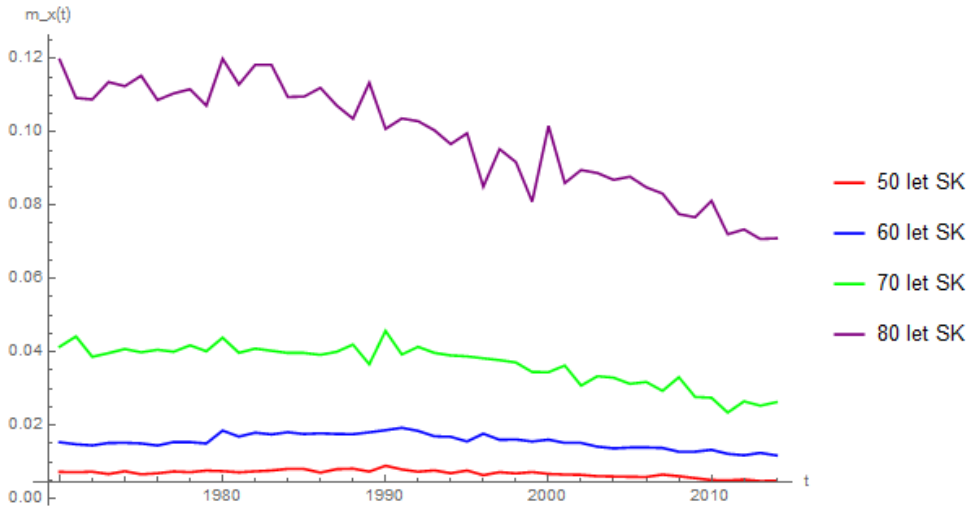
Před samotným modelováním úmrtnosti ve slovenské populaci uveďme opět nejprve obrázek 4.4 znázorňující vývoj specifické míry úmrtnosti v letech 1970, 1990 a 2010. Vidíme, že v úmrtnosti se pro jednotlivé roky vyskytují podobné rysy jako v případě české populace (vyšší úmrtnost v prvním roce života, "accident hump", exponenciální nárůst specifické míry úmrtnosti po třicátém roce života) a rovněž je z obrázku zřejmé snížení úmrtnosti zejména mezi lety 1990 a 2010, toto snížení se však nezdá být tolik výrazné jako v případě české populace. Vývoj specifické míry úmrtnosti ve věku 50, 60, 70 a 80 let mezi lety 1970 a 2014 lze posoudit v obrázku 4.5.

Pokles úmrtnosti ve věku 60 a 70 let ve sledovaném období v České a Slovenské republice je možné srovnat na základě obrázku 4.6. Pro oba věky nejprve specifická míra úmrtnosti české populace převyšuje míru úmrtnosti slovenské populace, v obou případech však úmrtnost kolísá okolo konstantní úrovně. Po roce 1985 začíná v obou populacích úmrtnost klesat, v té české však výrazně rychleji, proto lze pozorovat, že od roku 1990 je specifická míra úmrtnosti ve sledovaných věcích v české populaci nižší.

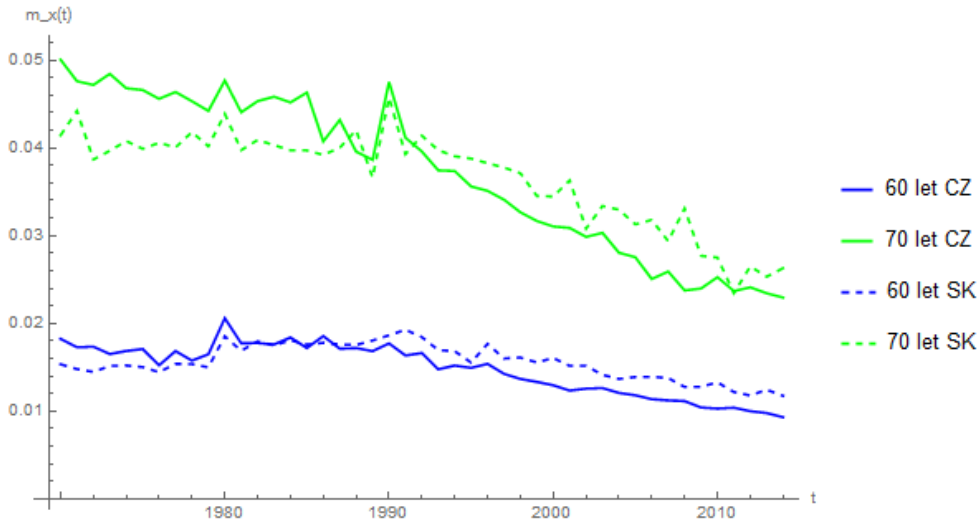
Stejně jako v případě české populace se omezme na modelování úmrtnosti ve věku $45 - 90$ let v letech $1970 - 2014$, opět využijme tří modelů popsaných v kapitole 2. Obrázek 4.7 znázorňuje standardizovaná Pearsonova rezidua Lee-Carterova, devianční rezidua Renshaw-Habermanova a Pearsonova rezidua Cairns-Blake--Dowdova modelu pro úmrtnost ve slovenské populaci pro věk 60 a 80 let. Ukazuje se, že pro modelování úmrtnosti je i ve slovenské populaci nejméně vhodný Carins-Blake-Dowdův model, ve kterém můžeme pozorovat rostoucí trend reziduí ve věku 60 let a klesající trend ve věku 80 let. Narozdíl od české populace ale ani rezidua Lee-Carterova modelu nevypadají jako náhodně rozprostřená okolo nuly. Rezidua Renshaw-Habermanova modelu se zdají být rozmístěna nejlépe.



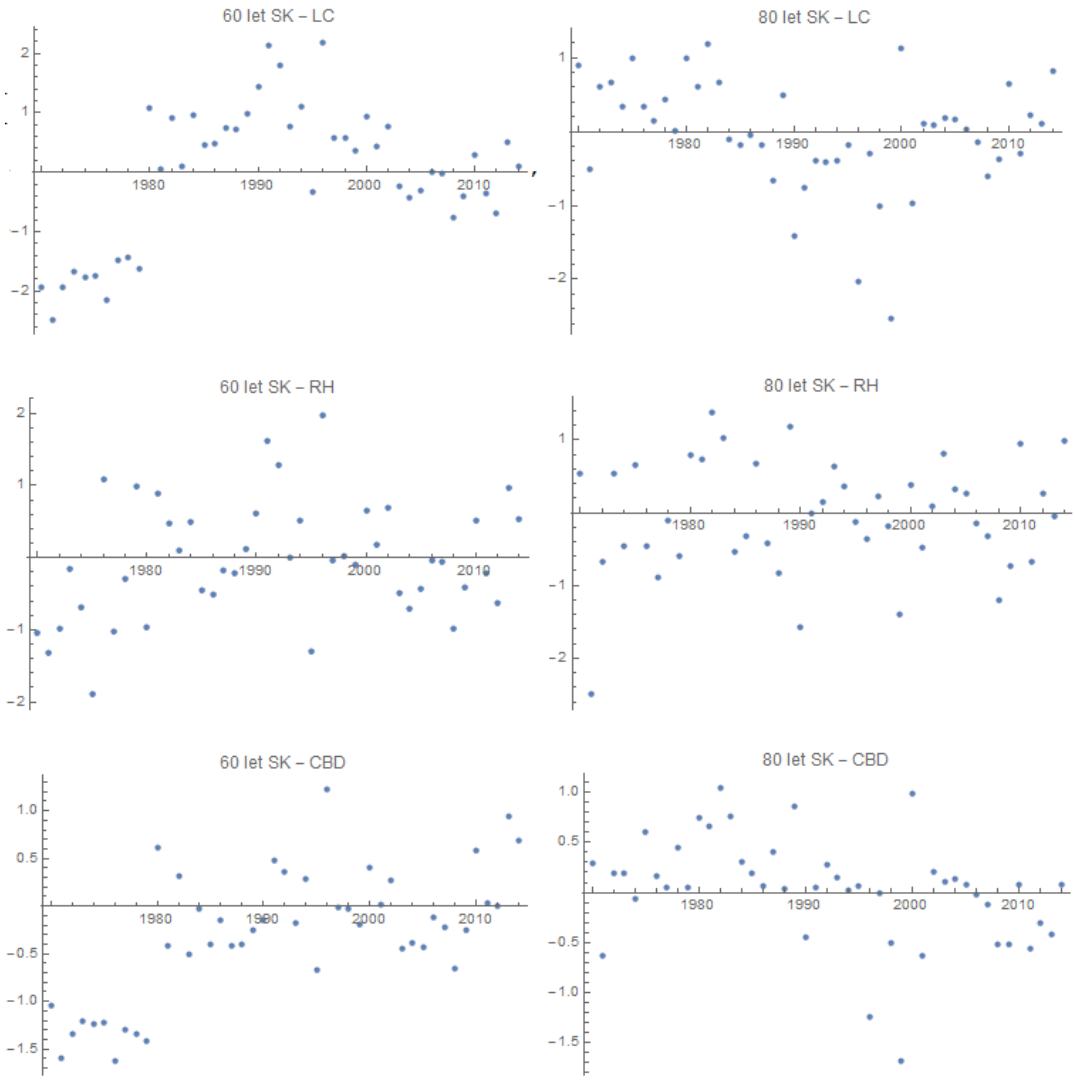
Obrázek 4.4: Specifická míra úmrtnosti v letech 1970, 1990 a 2010, Slovensko



Obrázek 4.5: Specifická míra úmrtnosti ve věku 50, 60, 70 a 80 let, Slovensko



Obrázek 4.6: Specifická míra úmrtnosti ve věku 60 a 70 let



Obrázek 4.7: Rezidua modelů úmrtnosti pro věk 60 a 80 let, Slovensko

Pro rozhodnutí o tom, který model je nejvhodnější pro modelování úmrtnosti slovenské populace, předpokládejme Poissonovo rozdělení počtů úmrtí $D_x(t)$ a vyberme model s nejvyšší hodnotou BIC. Hodnotu informačního kritéria spolu s maximem logaritmické věrohodnostní funkce pro každý z uvažovaných modelů uvádí tabulka 4.2. Nejvyššího BIC nabývá, stejně jako v případě české populace, Renshaw-Habermanův model. Jako nejméně vhodný se opět ukazuje model CBD.

Pokud bychom testovali vhodnost využití Lee-Carterova modelu jako podmodelu Renshaw-Habermanova modelu pomocí *likelihood-ratio* testu, zjistíme, že Renshaw-Habermanův model opět popisuje data výrazně lépe (testová statistika je rovna $3\,194{,}2 \gg 163{,}12 = \chi^2_{135}(0{,}95)$).

Model	**Maximální věrohodnost**	**Počet parametrů**	**BIC**	**Pořadí**
Lee-Carter	$-11\,681{,}3$	135	$-12\,196{,}7$	2
Renshaw-Haberman	$-10\,084{,}2$	270	$-11\,126{,}4$	1
Cairns-Blake-Dowd	$-14\,726{,}7$	90	$-15\,070{,}3$	3

Tabulka 4.2: Odhad parametrů metodou maximální věrohodnosti – Slovensko

4.3 Společné modelování úmrtnosti české a slovenské populace

V této podkapitole se zaměřme na současné modelování úmrtnosti v české a slovenské populaci pomocí zobecnění Lee-Carterova, Renshaw-Habermanova a Cairns-Blake-Dowdova modelu, která byla teoreticky popsána v kapitole 3.

Obdobně jako v případě modelování úmrtnosti v samostatných populacích vyberme nejvhodnější model, za předpokladu poissonovsky rozdělených počtů úmrtí, na základě hodnoty Bayesova informačního kritéria definovaného vztahem 4.1, povšimněme si, že vzhledem k dvojnásobnému počtu pozorovaných úmrtností ($N = 4\,140$) je počet parametrů v modelu penalizován více než v případě výše uvažovaných samostatných modelů.

Symboly příslušející české populaci označme horním indexem 1, tj. $m_x^{[1]}(t)$, $D_x^{[1]}(t)$, $E_x^{[1]}(t)$, $a_1^{[1]}(x)$, $a_2^{[1]}(x)$, $a_3^{[1]}(x)$, $a^{[1]}(x)$, $k_1^{[1]}(t)$, $k_2^{[1]}(t)$, $k^{[1]}(t)$ a $c^{[1]}(t-x)$. Pro označení slovenské populace využijme horní index 2.

4.3.1 Lee-Carterův model

Jako zobecnění Lee-Carterova modelu pro modelování úmrtnosti ve dvou populacích uvažujme modely, které byly v kapitole 3 pojmenovány jako:

- nezávislý model,

- model se společným K_1,
- kointegrovaný model,
- model společných faktorů,
- rozšířený model společných faktorů a
- model se společným A_2.

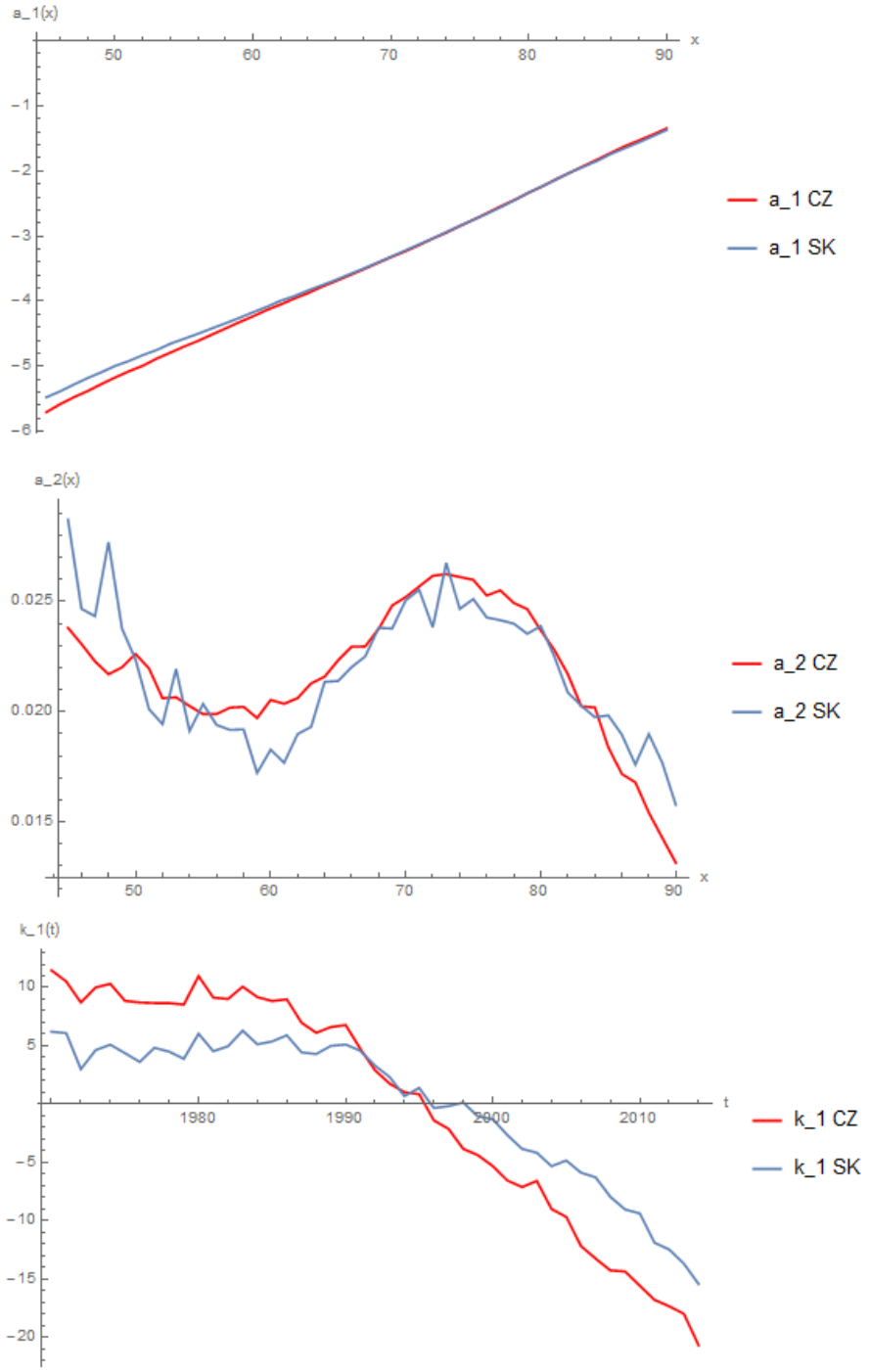
Obrázek 4.8 zobrazuje odhadnuté parametry Lee-Carterova modelu samostatně pro českou a slovenskou populaci, tj. parametry *nezávislého modelu*. Zaměříme-li se na graf časových řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$, resp. *k_1 CZ* a *k_1 SK*, které reprezentují úroveň úmrtnosti v čase t v jednotlivých populacích, vidíme, že do roku 1985 mají křivky v obou populacích přibližně konstantní trend. Po roce 1985 začínají obě úrovně úmrtnosti klesat a vidíme, že křivka vztahující se k české populaci, tj. $k_1^{[1]}(t)$, klesá výrazně rychleji než $k_1^{[2]}(t)$, a jelikož její hodnoty byly na počátku sledovaného období vyšší než v případě slovenské populace, dochází okolo roku 1995 dokonce k překřížení těchto křivek.

Některá uvažovaná zobecnění Lee-Carterova modelu předpokládají spojení časových řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ do řady společné pro obě populace $K_1(t)$. Podle údajů uvedených v tabulce 4.3 toto propojení však významně sníží maximum logaritmické věrohodnostní funkce i hodnotu informačního kritéria. Výjimkou je *rozšířený model společných faktorů*, který sice uvažuje spojení řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ do $K_1(t)$, avšak toto spojení kompenzuje přidáním dalších časových řad $(k^{[i]}(t))$ specifických pro každou populaci. Dle hodnoty Bayesova informačního kritéria můžeme toto zobecnění Lee-Carterova modelu pro společné modelování úmrtnosti v České a Slovenské republice považovat za nejvhodnější i přes nejvyšší počet parametrů. Druhá nejvyšší hodnota informačního kritéria přísluší *modelu se společným* A_2.

Tabulka 4.3 neobsahuje údaje pro *kointegrovaný model*, jeho odhady parametrů, stejně tak jako maximum logaritmické věrohodnostní funkce a hodnota BIC, jsou totiž shodné s těmi z *nezávislého modelu*, liší se však pohled na procesy $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ a tedy jejich případné predikované hodnoty pro $t' > t_n$.

Testujme možnost využití *kointegrovaného modelu* pomocí algoritmu uvedeného na straně 24. V prvním kroku ověřme nestacionaritu řad $k_1^{[i]}(t)$, $i = 1, 2$, z *nezávislého modelu* pomocí rozšířeného Dickey-Fullerova testu (pro zpoždění do pátého řádu) na hladině $\alpha = 0{,}05$, pro uvažovaná data získáme p-hodnotu rovnou 0,958 v případě české populace a 0,890 v případě slovenské. Nezamítáme tedy hypotézu, že $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ jsou nestacionární časové řady, a je proto možné pokračovat druhým krokem algoritmu, ve kterém odhadneme lineární regresní model pro $k_1^{[1]}(t)$:

$$k_1^{[1]}(t) = -5{,}30 \cdot 10^{-16} + 1{,}58 \cdot k_1^{[2]}(t) + u(t),$$



Obrázek 4.8: Parametry LC modelu pro českou a slovenskou populaci

kde $u(t)$ označuje náhodné odchylky. Dále opět pomocí rozšířeného Dickey-Fullerova testu testujme hypotézu o nestacionaritě řady $u(t)$ (rovněž pro 5 zpoždění), v případě této řady je p-hodnota testu rovna 0,118. Na hladině $\alpha = 0{,}05$ tedy nezamítáme hypotézu, že je řada nestacionární, a dle doporučeného algoritmu nemůžeme považovat procesy $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ za kointegrované.

Model	**Maximální věrohodnost**	**Počet parametrů**	**BIC**	**Pořadí**
Nezávislý model	−24 602,4	270	−25 726,7	3
Model se společným K_1	−28 977,9	226	−29 919,0	4
Model společných faktorů	−30 848,5	181	−31 602,2	5
Rozšířený model společných faktorů	−23 451,6	363	−24 963,2	1
Model se společným A_2	−24 659,9	225	−25 596,9	2

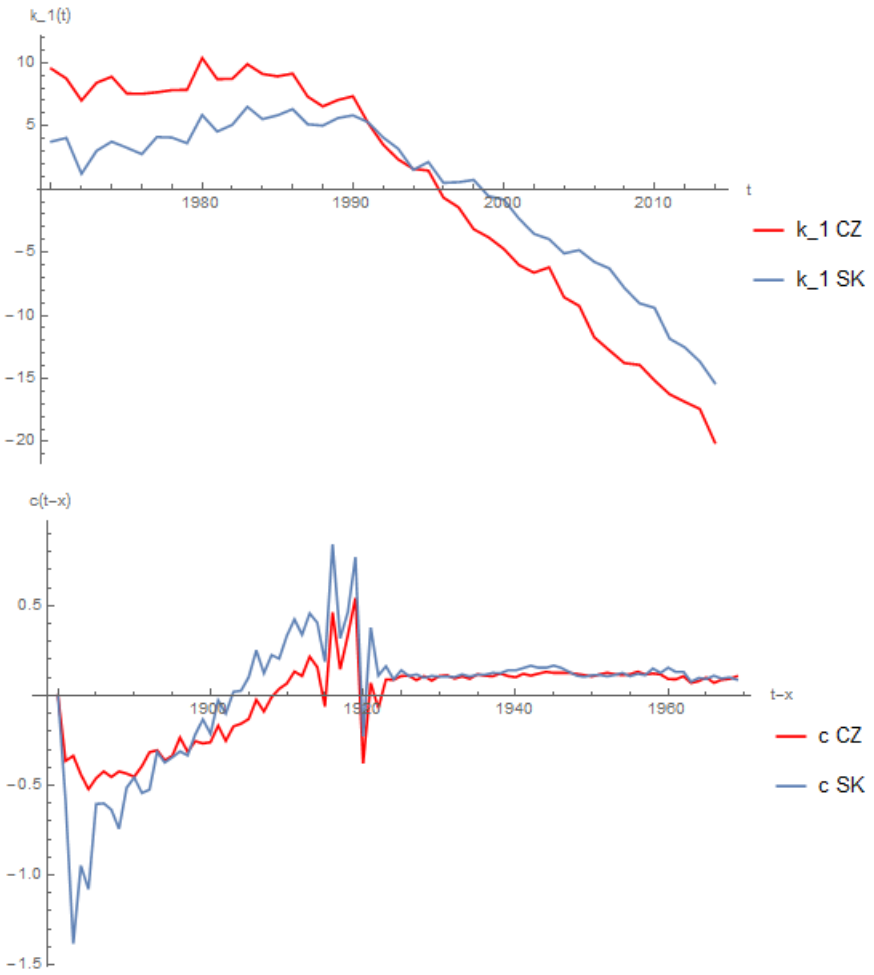
Tabulka 4.3: Odhad parametrů zobecněných LC modelů metodou maximální věrohodnosti

4.3.2 Renshaw-Habermanův model

Pro modelování úmrtnosti v samostatné české i slovenské populaci se ukázal Renshaw-Habermanův model jako nejvhodnější ze tří uvažovaných modelů. Věnujme se dále jeho modifikacím pro modelování úmrtnosti v obou populacích současně a zkoumejme, zda je možné některé časové řady v modelu pro obě populace sjednotit, aniž bychom snížili hodnotu Bayesova informačního kritéria. Pro všechna zobecnění uvažujme jednotné parametry $a_2^{[1]}(x) = a_2^{[2]}(x) =: A_2(x)$ a $a_3^{[1]}(x) = a_3^{[2]}(x) =: A_3(x)$, které dle [10] zaručí, že bude možné projektovat specifickou míru úmrtnosti v populacích tak, aby její podíl nedivergoval.

Obrázek číslo 4.9 znázorňuje vývoj parametrů $k_1^{[i]}(t)$ a $c^{[i]}(t-x)$ Renshaw-Habermanova modelu pro českou a slovenskou populaci (se společnými parametry $A_2(x)$ a $A_3(x)$), jejichž možné sjednocení budeme dále vyhodnocovat. Ve vývoji úrovní úmrtnosti $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ lze vidět obdobné tendence jako v případě těchto časových řad v Lee-Carterově modelu. Časové řady $c^{[i]}(t-x)$ zastupující efekt roku narození (kohortní efekt) vykazují v obou populacích podobný trend – v úmrtnosti osob narozených mezi lety 1885 a 1915 můžeme pozorovat rostoucí trend, v období okolo první světové války kolísání hodnot a následně přibližně konstantní trend (na podobné úrovni) v období od roku 1925. Obě časové řady se v české a slovenské populaci příliš neliší, můžeme tedy usuzovat, že je vhodné uvažovat modely, ve kterých jsou efekty kalendářního roku nebo roku narození modelovány společnou časovou řadou.

V tabulce 4.4 jsou uvedené uvažované modely pro zobecnění Renshaw-Habermanova modelu k modelování úmrtnosti ve více populacích z podkapitoly 3.2. Vidíme, že model, který uvažuje společnou řadu pro kohortní efekt $C(t-x) :=$



Obrázek 4.9: Parametry RH modelu pro českou a slovenskou populaci

$c^{[1]}(t-x) = c^{[2]}(t-x)$ dosahuje jen o několik set nižšího maxima logaritmické věrohodnostní funkce ve srovnání s *nezávislým modelem*. Díky počtu parametrů zahrnutých v těchto modelech a následné penalizaci za něj jsou rozdíly v Bayesově informačním kritériu velmi malé a jako nejvhodnější model na základě hodnoty tohoto kritéria můžeme vyhodnotit *model společné kohorty*.

Ve srovnání s Lee-Carterovým modelem nabývají téměř všechna uvažovaná zobecnění Renshaw-Habermanova modelu, mimo *modelu společných faktorů*, vyšší hodnoty Bayesova informačního kritéria.

Model	**Maximální věrohodnost**	**Počet parametrů**	**BIC**	**Pořadí**
Nezávislý model	−21 389,4	450	−23 263,3	2
Model se společným K_1	−22 271,2	405	−23 957,7	3
Model společné kohorty	−21 685,7	361	−23 189,0	1
Model společných faktorů	−27 680,2	316	−28 996,1	4

Tabulka 4.4: Odhad parametrů zobecněných RH modelů metodou maximální věrohodnosti

4.3.3 Cairns-Blake-Dowdův model

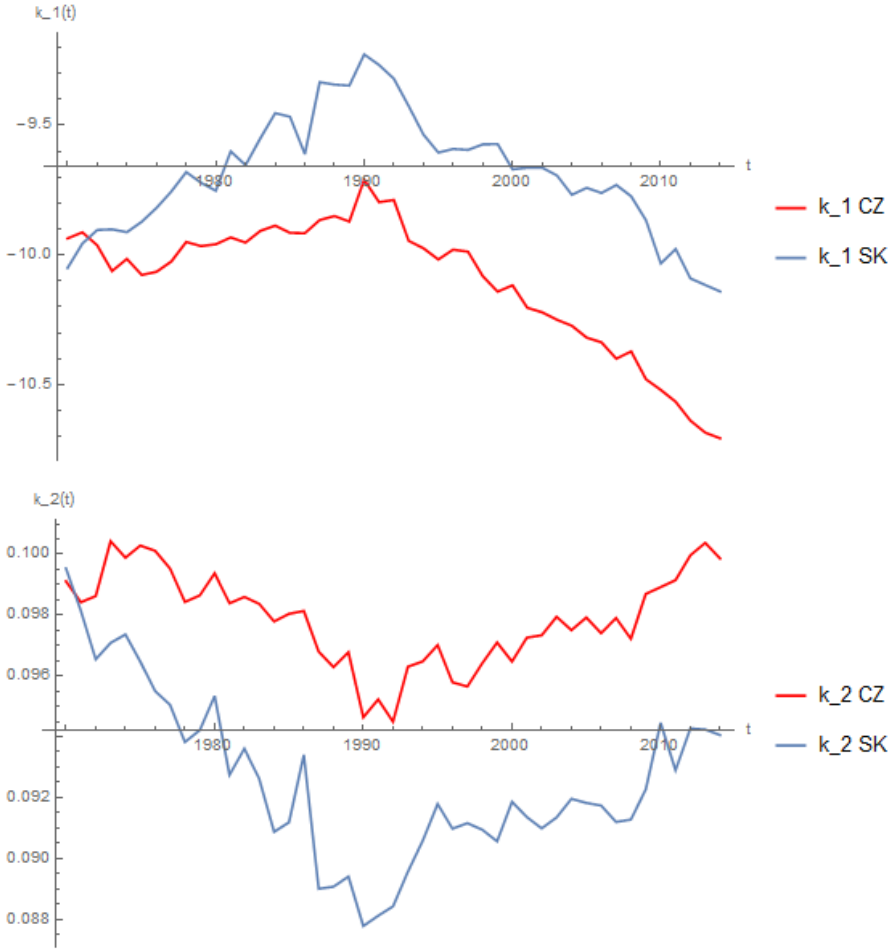
Přestože se Cairns-Blake-Dowdův model ukázal jako nejméně vhodný pro modelování úmrtnosti v české a ve slovenské populaci, zaměřme se v následující části i na zobecnění tohoto modelu pro současné modelování úmrtnosti ve dvou populacích.

Obrázek číslo 4.10 zachycuje časové řady $k_1^{[i]}(t)$ a $k_2^{[i]}(t)$ pro jednotlivé populace, tedy parametry *nezávislého modelu*. Vidíme, že odhady mezi jednotlivými populacemi se značně liší, můžeme proto očekávat, že spojení některé z časových řad do společné řady $K_1(t)$ nebo $K_2(t)$ se ukáže jako nevhodné.

V tabulce 4.5 jsou zaznamenána maxima logaritmické věrohodnostní funkce modelů úmrtnosti ve dvou populacích a jim příslušné hodnoty Bayesova informačního kritéria. Vidíme, že modely se spojenou některou z časových řad se pro dané populace ukazují jako významně horší než *nezávislý model*.

4.4 Předpověď úmrtnosti české a slovenské populace

Jelikož je Cairns-Blake-Dowdův model nejméně vhodným i mezi nezávislými modely (respektive mezi modely pro samostatné modelování úmrtnosti v jednotlivých populacích) a spojení časových řad vedla k významnému snížení maxim



Obrázek 4.10: Parametry CBD modelu pro českou a slovenskou populaci

Model	**Maximální věrohodnost**	**Počet parametrů**	**BIC**	**Pořadí**
Nezávislý model	−33 049,1	180	−33 798,7	1
Model se společným K_1	−186 837,0	135	−187 399,2	3
Model se společným K_2	−157 334,0	135	−157 896,2	2

Tabulka 4.5: Odhad parametrů zobecněných CBD modelů metodou maximální věrohodnosti

logaritmické věrohodnostní funkce, uvažujme pro predikci úmrtnosti české a slovenské populace pouze varianty Lee-Carterova a Renshaw-Habermanova modelu pro současné modelování úmrtnosti v obou populacích, které dosahovaly nejvyšší hodnoty Bayesova informačního kritéria. Zaměřme se tedy dále na *rozšířený model společných faktorů* jako zobecnění Lee-Carterova modelu a *model společné kohorty* jako variantu Renshaw-Habermanova modelu.

4.4.1 Rozšířený Lee-Carterův model společných faktorů

Jak bylo uvedeno v podkapitole 4.3.1, můžeme u časových řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ reprezentujících úroveň úmrtnosti v modelu pro samostatné populace (tj. v nezávislém modelu) pozorovat změnu trendu – přibližně do roku 1985 je konstantní a následně začíná úroveň úmrtnosti významně klesat. Obdobně je tomu i v případě spojené časové řady $K_1(t)$ v rozšířeném modelu společných faktorů, jak je možné vidět dále z obrázku 4.11. Zaměřme se proto při odhadu vhodného modelu pro projekci časových řad na období po roce 1985.

Uvažujme nejprve modely pro samostatné populace. Jak je navrženo v části 2.1.2, pro modelování a předpovídání časových složek Lee-Carterových modelů je možné využít náhodné procházky s driftem.
V případě české populace byla hodnota parametru driftu odhadnuta (pomocí softwaru EViews) jako $-0{,}996$, model pro časovou řadu $k_1^{[1]}(t)$ je tedy možné zapsat ve tvaru

$$k_1^{[1]}(t) - k_1^{[1]}(t-1) = -0{,}996 + \varepsilon(t).$$

Pokud testujeme hypotézu o nestacionaritě reziduální složky $\{\varepsilon(t)\}$ pomocí rozšířeného Dickey-Fullerova testu (pro pátý řád zpoždění), získáme p-hodnotu 0,000, hypotézu je tedy dle výsledků tohoto testu možné zamítnout a rezidua lze považovat za stacionární.

Dále testujme, zda mají rezidua nulovou střední hodnotu, jsou nekorelovaná, homoskedastická a normálně rozdělená. Nulová střední hodnota byla testována v EViews pomocí t-testu. Pro otestování hypotézy, že rezidua nejsou korelovaná, byl využit Ljung-Boxův test (testující korelaci do pátého řádu), bližší informace o tomto testu lze nalézt v [12]. Homoskedasticita reziduální složky byla testována pomocí testu ARCH, rovněž viz [12]. Pro testování normality reziduální složky byl využit Shapiro-Wilkův test v Mathematice.
p-hodnoty jednotlivých testů jsou

t-test:	1,000,
Ljung-Box:	0,465,
ARCH:	0,533,
Shapiro-Wilk:	0,469,

na hladině spolehlivosti 5 % tedy nezamítáme nulovou střední hodnotu, nekorelovanost, homoskedasticitu ani normalitu reziduální složky. Pro českou populaci se tedy náhodná procházka s driftem zdá být vhodným modelem pro časovou řadu

$k_1^{[1]}(t)$ a není tedy třeba uvažovat pro modelování časové složky Lee-Carterova modelu $ARIMA$ proces s nenulovými řády p a q, případně jinou diferencí.

Ve slovenské populaci uvažujme rovněž pro modelování úrovně úmrtnosti v kalendářním roce náhodnou procházku s driftem, pro slovenskou populaci ji lze zapsat jako

$$k_1^{[2]}(t) - k_1^{[2]}(t-1) = -0{,}686 + \varepsilon(t).$$

I v případě slovenské populace je p-hodnota rozšířeného Dickey-Fullerova testu pro reziduální složku modelu rovna 0,000, rezidua tedy lze považovat za stacionární.
V případě testů dalších vlastností reziduí jsou p-hodnoty rovny

t-test:	1,000,
Ljung-Box:	0,388,
ARCH:	0,218,
Shapiro-Wilk:	0,282

a tedy opět na hladině spolehlivosti 5 % nezamítáme nulovou střední hodnotu, nekorelovanost, homoskedasticitu ani jejich normalitu. Ve slovenské populaci lze tedy pro modelování časové řady $k_1^{[1]}(t)$ rovněž využít náhodnou procházku s driftem.

Pro projekci úmrtnosti pomocí rozšířeného modelu společných faktorů modelujme spojený efekt kalendářního roku $K_1(t)$ náhodnou procházkou s driftem a samostatné efekty $k^{[i]}(t)$ procesem $AR(1)$ tak, jak navrhuje [9]. Po odhadnutí parametrů lze modely pro jednotlivé řady zapsat jako

$$\begin{aligned} K_1(t) - K_1(t-1) &= -0{,}893 + \varepsilon(t), \\ k_1^{[1]}(t) &= 0{,}026 + 0{,}982 \cdot k_1^{[1]}(t-1) + \varepsilon^{[1]}(t), \\ k_1^{[2]}(t) &= 0{,}030 + 0{,}955 \cdot k_1^{[2]}(t-1) + \varepsilon^{[2]}(t). \end{aligned}$$

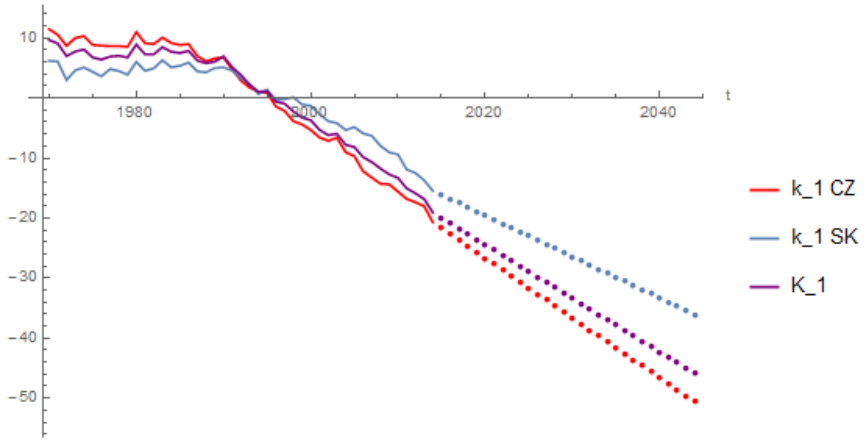
Testujeme-li stacionaritu reziduálních složek v modelech pomocí rozšířeného Dickey-Fullerova testu, získáme ve všech případech p-hodnotu nižší než 0,05 (0,001 v případě $\{\varepsilon(t)\}$, 0,021 u reziduí modelu efektu kalendářního roku české populace $\{\varepsilon^{[1]}(t)\}$ a 0,039 v případě reziduí slovenské populace $\{\varepsilon^{[2]}(t)\}$), všechny reziduální složky tedy můžeme považovat za stacionární.
Tabulka 4.6 dále zachycuje p-hodnoty testů nulové střední hodnoty, nekorelovanosti, homoskedasticity a normality reziduálních složek. Vidíme, že mimo předpokladu normality v případě reziduí spojeného efektu kalendářního roku a mimo předpokladu nekorelovanosti v případě specifické řady pro českou populaci, testované hypotézy o vlastnostech reziduálních složek nezamítáme.

Obrázek 4.11 znázorňuje projekci efektů kalendářního roku v české populaci a ve slovenské populaci, tj. řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$, a dále projekci spojeného efektu pro obě populace v rozšířeném modelu společných faktorů. Populačně specifické

řada	t-test	Ljung-Box	ARCH	Shapiro-Wilk
$K_1(t)$	1,000	0,582	0,677	0,010
$k^{[1]}(t)$	1,000	0,012	0,247	0,345
$k^{[2]}(t)$	1,000	0,696	0,625	0,115

Tabulka 4.6: p-hodnoty testů reziduálních složek modelů časových řad v modelu LC

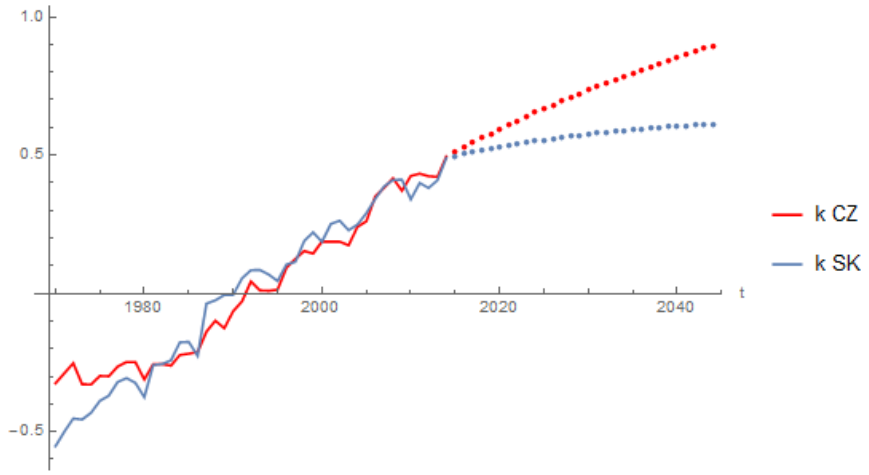
řady $k^{[1]}(t)$ a $k^{[2]}(t)$ z tohoto modelu znázorňuje spolu s projektovanými hodnotami pomocí modelů $AR(1)$ obrázek 4.12.



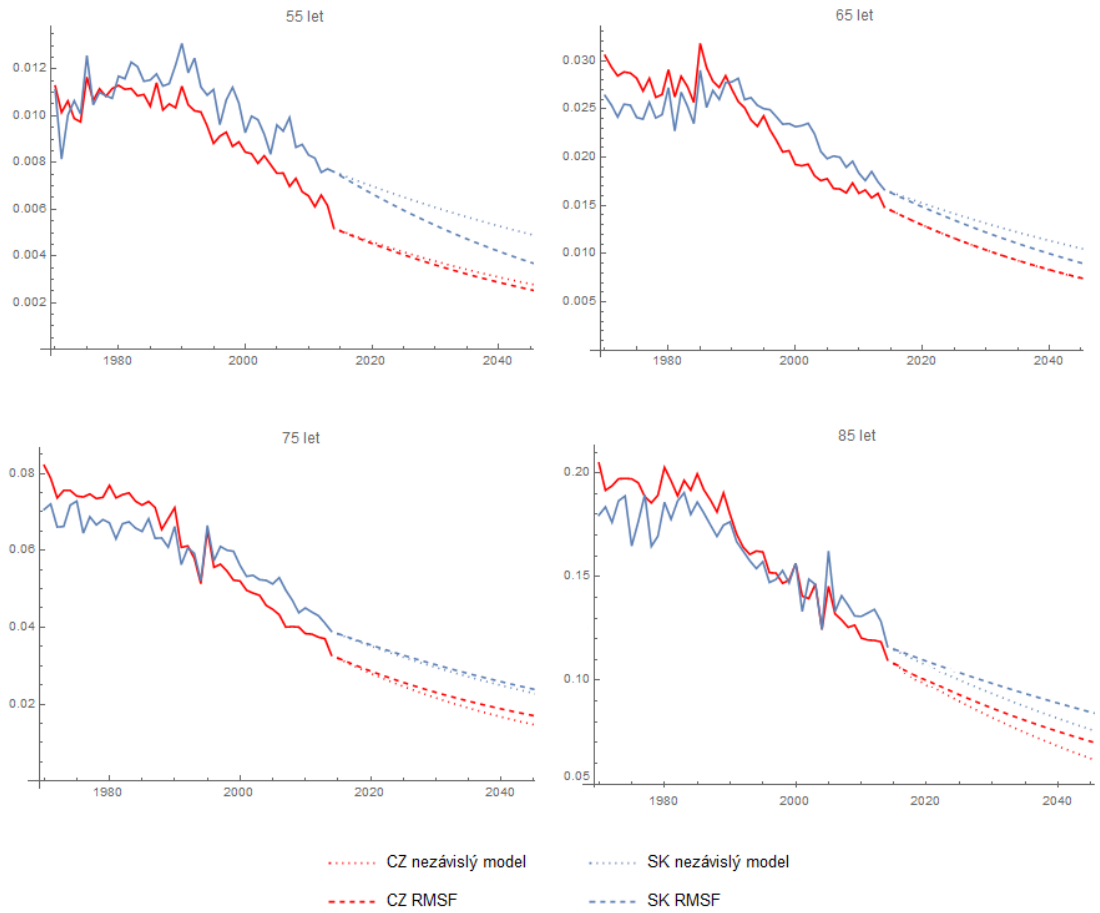
Obrázek 4.11: Projekce efektů kalendářních roků LC modelu

Obrázek 4.13 znázorňuje projekce specifické míry úmrtnosti v české a slovenské populaci ve věku 55, 65, 75 a 85 let, a to pomocí Lee-Carterových modelů pro samostatné populace i pomocí rozšířeného modelu společných faktorů (RMSF). Vidíme, že v případě využití rozšířeného modelu společných faktorů pozorujeme rychlejší pokles míry úmrtnosti v nižších věcích uvažovaného věkového rozsahu (především v případě úmrtnosti ve věku 55 let), zatímco ve vyšších věcích (75 a 85 let) dochází v případě rozšířeného modelu společných faktorů k výrazně pomalejšímu poklesu věkově specifické míry úmrtnosti v porovnání s modely pro samostatné populace.

Dále je možné porovnat úmrtnosti predikované pomocí jednotlivých modelů v tabulce 4.7, která uvádí srovnání pravděpodobností, že osoba, která je v roce 2015 ve věku 45, 55 a 65 let, bude žít ještě 20 let. V tabulce jsou uvedené tyto pravděpodobnosti pro českou i slovenskou populaci predikované podle obou uvažovaných modelů. Obdobně jako z grafického znázornění 4.13 můžeme konstatovat, že v případě mladších věků je rozšířeným modelem společných faktorů predikován rychlejší pokles míry úmrtnosti, ve věku 45 let je tedy pravděpodobnost přežití dvaceti let vyšší než v případě modelů samostatných populací. Naopak v případě osob ve vstupním věku 65 let dochází při využití společného modelu ke snížení pravděpodobnosti přežití.



Obrázek 4.12: Projekce specifických efektů v rozšířeném modelu společných faktorů



Obrázek 4.13: Projekce specifické míry úmrtnosti LC

věk v 2015	CZ nezávislý	CZ RMSF	SK nezávislý	SK RMSF
45	0,910	0,912	0,875	0,888
55	0,796	0,791	0,745	0,750
65	0,547	0,528	0,469	0,455

Tabulka 4.7: Dvacetileté pravděpodobnosti přežití v LC modelu

4.4.2 Renshaw-Habermanův model společné kohorty

Časové řady reprezentující efekt kalendářního roku v Renshaw-Habermanově modelu (pro samostatné populace i v modelu pro obě populace se společnou řadou kohortního efektu) vykazují podobný trend jako v případě Lee-Carterova modelu. Zhruba do roku 1985 tedy můžeme pozorovat přibližně konstantní trend a následně postupný pokles úrovně úmrtnosti. Opět se proto při odhadu vhodného modelu pro projekci časových řad zachycujících efekt kalendářního roku zaměřme na období až po roce 1985.

Řady $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ v samostatných populacích, respektive v nezávislém modelu, modelujme opět pomocí náhodné procházky s driftem

$$\begin{aligned} k_1^{[1]}(t) - k_1^{[1]}(t-1) &= -0{,}989 + \varepsilon^{[1]}(t), \\ k_1^{[2]}(t) - k_1^{[2]}(t-1) &= -0{,}694 + \varepsilon^{[2]}(t), \end{aligned}$$

v případě obou procesů zamítáme na hladině spolehlivosti 5 % nestacionaritu reziduální složky pomocí rozšířeného Dickey-Fullerova testu, p-hodnota testu je v případě české populace rovna 0,011 a v případě slovenské populace 0,040, obě reziduální složky tedy lze považovat za stacionární. Pokud dále testujeme hypotézy o nulové střední hodnotě, nekorelovanosti, homoskedasticitě a normalitě reziduí, vidíme z tabulky 4.8, že všechny p-hodnoty jsou vyšší než 0,05, tedy žádnou z uvedených hypotéz nezamítáme a uvedené náhodné procházky s driftem můžeme považovat za vhodné pro modelování časových řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$.

řada	t-test	Ljung-Box	ARCH	Shapiro-Wilk
$k^{[1]}(t)$	1,000	0,444	0,582	0,436
$k^{[2]}(t)$	1,000	0,350	0,638	0,998

Tabulka 4.8: p-hodnoty testů reziduálních složek modelů $k_1^{[i]}(t)$ v RH modelu

K modelování řad $k_1^{[1]}(t)$ a $k_1^{[2]}(t)$ ve společném modelu pro obě populace, který obsahuje spojenou řadu pro kohortní efekt, využijme dle podkapitoly 3.5.1 model $RWAR$ – tedy model, ve kterém efekt kalendářního roku v české populaci modelujeme pomocí náhodné procházky s driftem a rozdíl v tomto efektu mezi českou a slovenskou populací pomocí autoregresního procesu prvního řádu.
Jelikož o vektorech náhodných odchylek $(Z_1^{[1]}(y), Z_1^{[2]}(y))$ v modelu $RWAR$ předpokládáme, že jsou vzájemně nezávislé s dvojrozměrným normálním rozdělením s nulovou střední hodnotou a konstantní kovarianční maticí, je možné odhadnout parametry modelu $\mu_1^{[1]}$, $\mu_1^{[2]}$ a $\phi_1^{[2]}$ a hodnoty v kovarianční matici pomocí metody

maximální věrohodnosti. Výraz, který maximalizujeme je možné dle [10] zapsat jako

$$\sum_{t=t_1}^{t_n-1} \log f_{N_2(\boldsymbol{\mu}+\boldsymbol{\phi}_{RWAR}\cdot\boldsymbol{S}_t,\,\boldsymbol{V})}(\boldsymbol{S}_{t+1}),$$

kde argumentem logaritmu je hustota dvojrozměrného normálního rozdělení s vektorem středních hodnot $\boldsymbol{\mu} + \boldsymbol{\phi}_{RWAR} \cdot \boldsymbol{S}_t$ a kovarianční maticí $\boldsymbol{V}$ v bodě $\boldsymbol{S}_{t+1}$, přičemž jednotlivé symboly označují:

$$\begin{aligned}
\boldsymbol{\mu} &= \begin{pmatrix} \mu_1^{[1]} \\ \mu_1^{[2]} \end{pmatrix} \\
\boldsymbol{\phi}_{RWAR} &= \begin{pmatrix} 1 & 0 \\ 0 & \phi_1^{[2]} \end{pmatrix} \\
\boldsymbol{S}_t &= \begin{pmatrix} k_1^{[1]}(t) \\ k_1^{[1]}(t) - k_1^{[2]}(t) \end{pmatrix}.
\end{aligned}$$

Po vypočítání odhadů parametrů $\mu_1^{[1]}$, $\mu_1^{[2]}$ a $\phi_1^{[2]}$ lze model $RWAR$ pro úmrtnost v české populaci a rozdíly v úmrtnosti české a slovenské populace zapsat jako

$$\begin{aligned}
k_1^{[1]}(t) - k_1^{[1]}(t-1) &= -0{,}972 + Z_1^{[1]}(t), \\
k_1^{[1]}(t) - k_1^{[2]}(t) &= -0{,}522 + 0{,}899 \cdot (k_1^{[1]}(t-1) - k_1^{[2]}(t-1)) + Z_1^{[2]}(t).
\end{aligned}$$

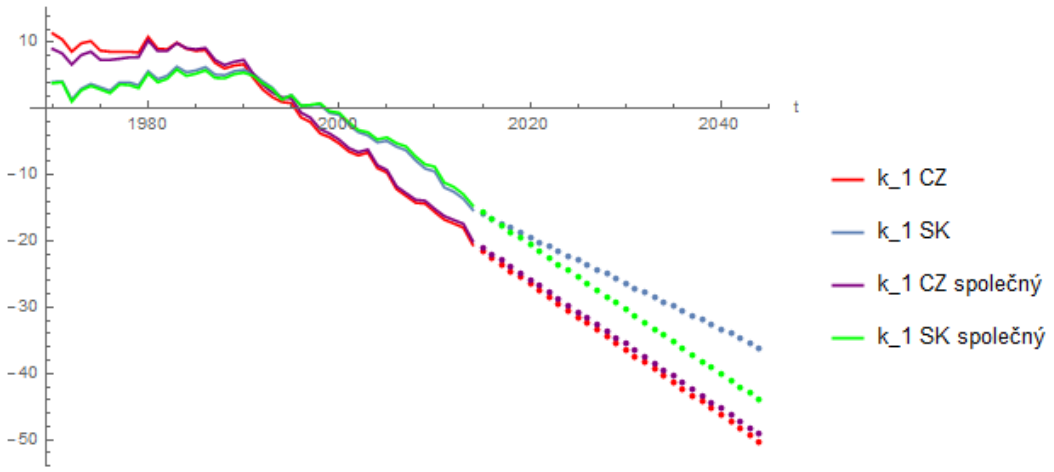
Kovarianční matice $\boldsymbol{V}$ má tvar

$$\begin{pmatrix} 0{,}708 & 0{,}424 \\ 0{,}424 & 0{,}608 \end{pmatrix}.$$

Testujeme-li hypotézu, že vektory náhodných odchylek pocházejí z dvojrozměrného normálního rozdělení s nulovou střední hodnotou a kovarianční maticí $\boldsymbol{V}$ pomocí Kolmogorovova-Smirnovova testu, získáme p-hodnotu rovnou 0,953, splnění předpokladu dvojrozměrného normálního rozdělení s konstantní kovarianční maticí tedy nezamítáme. Pro testování, zda jsou vektory náhodných odchylek vzájemně nekorelované, byl využit vícerozměrný Ljung-Boxův test v softwaru R, pro detaily viz [16], p-hodnota pro zpoždění do pátého řádu je rovna 0.300, nezamítáme tedy hypotézu nekorelovanosti, a jelikož jsme nezamítali ani hypotézu o dvojrozměrném normálním rozdělení, můžeme vektory náhodných odchylek považovat i za nezávislé. Předpoklady modelu jsou tedy splněny.

Srovnání projekcí efektů kalendářních roků v české a slovenské populaci v Renshaw-Habermanově modelu pro samostatné populace a v zobecněném modelu se spojeným kohortním efektem (a se společnými parametry $A_2(x)$ a $A_3(x)$) je možné v obrázku 4.14. Vidíme, že v případě společného modelování tohoto efektu dochází v české populaci k mírnému zpomalení poklesu projektované úrovně úmrtnosti v porovnání s projekcí tohoto efektu v modelu pro samostatnou českou populaci. Naopak v případě slovenské populace je pokles úrovně úmrtnosti ve společném modelu se spojeným kohortním efektem významně rychlejší.

Pro modelování kohortních efektů v samostatných modelech i modelování spojeného kohortního efektu ve společném modelu vyberme model $ARIMA(p, d, q)$,



Obrázek 4.14: Projekce efektu kalendářního roku v RH modelu

který dosahuje nejvyšší hodnoty Bayesova informačního kritéria z $ARIMA$ procesů s parametry p, d, $q \in \{0, 1, 2\}$. Jelikož ve všech řadách kohortních efektů můžeme pozorovat velké výkyvy v letech $1915 - 1925$, zaměřme se při modelování tohoto efektu na období po roce 1925.

Pro modelování kohortního efektu v české populaci se ukazuje jako nejvhodnější proces $ARIMA(2,0,2)$. Ve slovenské populaci dosahuje nejvyšší hodnoty Bayesova informačního kritéria náhodná procházka s driftem, $ARIMA(0,1,0)$, a pro modelování spojeného kohortního efektu je nejlepším modelem $ARIMA(1,0,2)$.

$$\begin{aligned}
c^{[1]}(y) &= 0.020 + 1.117 \cdot c^{[1]}(y-1) - 0.357 \cdot c^{[1]}(y-2) \\
&+ \varepsilon^{[1]}(y) - 0.910 \cdot \varepsilon^{[1]}(y-1) + 0.996 \cdot \varepsilon^{[1]}(y-2), \\
c^{[2]}(y) &= c^{[2]}(y-1) - 0.002 + \varepsilon^{[2]}(y), \\
C(y) &= 0.139 + 0.853 \cdot C(y-1) \\
&+ \varepsilon(y) - 0.416 \cdot \varepsilon(y-1) + 0.469 \cdot \varepsilon(y-2).
\end{aligned}$$

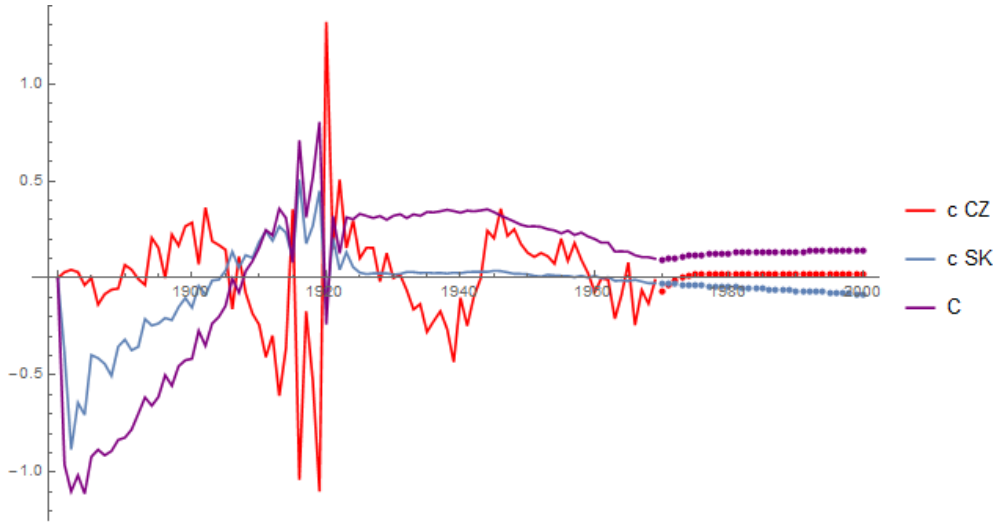
Testujeme-li stacionaritu reziduálních složek pomocí rozšířeného Dickey-Fullerova testu, získáme ve všech třech případech p-hodnotu rovnou 0,000, z tohoto důvodu zamítáme nulové hypotézy, které říkají, že reziduální složky nejsou stacionární. p-hodnoty testů nulové střední hodnoty, nekorelovanosti, homoskedasticity a normality jednotlivých reziduálních složek uvádí tabulka 4.9. Vidíme, že na hladině spolehlivosti 5 % není splněn pouze předpoklad normality u reziduí kohortních efektů pro samostatnou českou a slovenskou populaci. Splnění ostatních předpokladů nezamítáme.

Obrázek 4.15 znázorňuje vývoj kohortních efektů v letech $1880 - 1969$ a následně uvádí predikce pro vývoj kohortních efektu mimo rozsah využitých dat. Vidíme, že predikované hodnoty pro jednotlivé kohortní efekty jsou téměř konstantní, kohortní efekt tedy pro projekci úmrtnosti osob narozených po roce 1969 v uvažovaných modelech není příliš významný.

Obdobně jako v případě Lee-Carterových modelů uveďme nyní i pro Renshaw-Habermanův model obrázek porovnávající vývoj specifických měr úmrtnosti ve

řada	t-test	Ljung-Box	ARCH	Shapiro-Wilk
$c^{[1]}(y)$	1,000	0,169	0,999	0,027
$c^{[2]}(y)$	1,000	0,127	0,952	0,006
$C(y)$	1,000	0,547	0,832	0,074

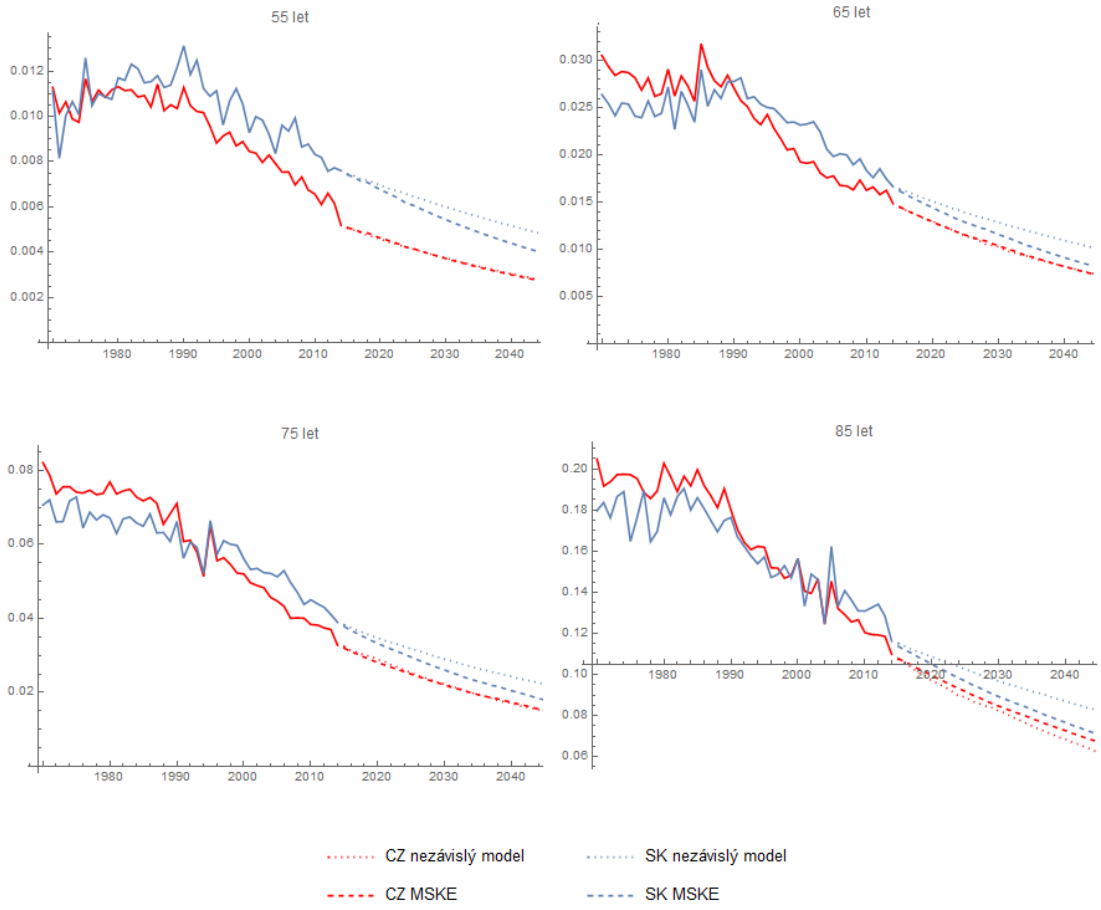
Tabulka 4.9: p-hodnoty testů reziduálních složek kohortních efektů RH modelu



Obrázek 4.15: Projekce kohortního efektu v RH modelu

věcích 55, 65, 75 a 85 let. Predikovaný vývoj specifických měr úmrtnosti dle samostatných modelů pro českou a slovenskou populaci a dle společného modelu se spojeným kohortním efektem (MSKE) je možné srovnat v obrázku 4.16.Vidíme, že v české populaci dochází v 55, 65 a 75 letech jen k nepatrné změně úmrtnosti při využití společného modelu, ve věku 85 let můžeme pozorovat výraznější zpomalení poklesu úmrtnosti. Naopak u slovenské populace je ve společném modelu ve všech uvažovaných věcích zřejmý rychlejší pokles úmrtnosti ve srovnání s modelem pro samostatnou slovenskou populaci.

Tabulka 4.10 uvádí pravděpodobnosti, že se osoby, které jsou v roce 2015 ve věku 45, 55 a 65 let dožijí věku o dvacet let vyššího. V tabulce je možné srovnat tyto pravděpodobnosti pro populace České i Slovenské republiky, a to podle samostatných modelů a společného modelu se spojeným kohortním efektem. V české populaci se pravděpodobnost přežití osob ve věku 45 a 55 let v obou modelech téměř neliší, k výraznějšímu snížení dvacetileté pravděpodobnosti přežití dochází pouze u osob v počátečním věku 65 let. V případě slovenské populace je změna opět výraznější, v případě všech vstupních věků se pravděpodobnosti přežití zvýšily, a to postupně o jednu, dvě a tři setiny pro osoby ve věku 45, 55 a 65 let v roce 2015.



Obrázek 4.16: Projekce specifické míry úmrtnosti RH

věk v 2015	CZ nezávislý	CZ MSKE	SK nezávislý	SK MSKE
45	0,911	0,910	0,877	0,887
55	0,798	0,796	0,750	0,768
65	0,543	0,538	0,466	0,496

Tabulka 4.10: Dvacetileté pravděpodobnosti přežití v RH modelu

Závěr

Cílem této diplomové práce bylo popsat možnosti rozšíření vybraných stochastických modelů úmrtnosti v samostatných populacích na modely úmrtnosti ve dvou či více korelovaných populacích a dále ilustrovat tyto postupy na mírách úmrtnosti v populacích České a Slovenské republiky.

V teoretické části byly nejprve definovány základní pojmy z oblasti demografie, které se užívají v souvislosti s úmrtnosti, následně byl popsán Lee-Carterův, Renshaw-Habermanův a Cairns-Blake-Dowdův model pro stochastické modelování úmrtností. Ve třetí kapitole teoretické části byly uvedeny možnosti pro rozšíření těchto modelů na současné modelování úmrtnosti ve více populacích, které zohledňují vztah mezi uvažovanými populacemi. Rovněž byly popsány metody pro projekci časových řad v těchto modelech, které zaručí splnění postačující podmínky pro konvergenci podílu věkově specifických měr úmrtnosti v těchto populacích.

V praktické části byly navržené postupy pro samostatné modelování úmrtnosti v dané populaci či společné modelování úmrtnosti ve více populacích ilustrovány na údajích o úmrtnosti v české a slovenské populaci v letech $1970-2014$ ve věcích $45-90$ let. Ukázalo se, že pro modelování úmrtnosti v samostatných populacích je v obou případech nejvhodnějším modelem Renshaw-Habermanův model, naopak nejméně vhodným modelem je Cairns-Blake-Dowdův model.
Dále byla uvažována zobecnění Lee-Carterova a Renshaw-Habermanova modelu pro současné modelování úmrtnosti v české a slovenské populaci. Na základě hodnoty Bayesova informačního kritéria byl jako nejvhodnější model vybrán *rozšířený Lee-Carterův model společných faktorů* a *Renshaw-Habermanův model společné kohorty*. Pro tyto modely byla projektována věkově specifická míra úmrtnosti, která byla porovnána s projektovanými hodnotami pomocí samostatných modelů pro jednotlivé populace. Při využití Lee-Carterova modelu došlo u *rozšířeného modelu společných faktorů* ke zrychlení poklesu specifické míry úmrtnosti v nižších věcích uvažovaného rozsahu (do 65 let), naopak ve vyšších věcích projektuje tento společný model pomalejší pokles specifické míry úmrtnosti ve srovnání se samostatnými modely pro jednotlivé populace. V případě *Renshaw-Habermanova modelu společné kohorty* klesá ve všech uvažovaných věcích projektovaná specifická míra úmrtnosti české populace pomaleji a specifická míra úmrtnosti slovenské populace rychleji než projektovaná míra úmrtnosti v samostatných modelech.

Seznam použité literatury

[1] Cipra, T. *Pojistná matematika - teorie a praxe.* 2. vydání. Ekopress, Praha 2006. ISBN 978-80-86929-11-6.

[2] Cipra, T. *Penze: kvantitativní přístup.* 2. vydání. Ekopress, Praha 2012. ISBN 978-80-86929-87-3.

[3] Pitacco, E., Denuit, M., Haberman, S., Olivieri, A. M. *Modelling Longevity Dynamics for Pensions and Annutiy Business.* Oxford University Press 2009. ISBN 978-0-19-954727-2.

[4] Český statistický úřad *Úmrtnostní tabulky - Metodické poznámky* [online]. [citováno 2. 3. 2016]. Dostupné z: https://www.czso.cz/csu/czso/umrtnostni-tabulky-metodika

[5] Cairns, A. J. G., Blake, D. et al. *A quantitative comparison of stochastic mortality models using data from England and Wales and the United States.* North American Actuarial Journal, Vol. 13, (2009), s. 1-35

[6] Lee, R. D., Carter, L. R. *Modelling and forecasting U. S. Mortality.* Journal of the American Statistical Association, Vol. 87, (1992), s. 659-671.

[7] Renshaw, A. E., Haberman, S. *A cohort-based extension to the Lee-Carter model for mortality reduction factors.* Insurance: Mathematics Economics, Vol. 38(3), (2006), s. 556-570.

[8] Cairns, A. J. G., Blake, D., Dowd, K. *A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration.* The Journal of Risk and Insurance, Vol. 73(4), (2006), s. 687-718.

[9] Li, N., Lee, R. D. *Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method.* Demography, Vol. 42(3), (2005), s. 575-594.

[10] Li, J. S., Zhou, R., Hardy, M. R. *A step-by-step guide to building two-population stochastic mortality models.* Insurance: Mathematics and Economics, Vol. 63, (2015), s. 121-134

[11] Li, J. S., Hardy, M. R. *Measuring basis risk in longevity hedges.*

[12] Startz, R. *EViews Illustrated for Version 9* [online]. [citováno 22. 4. 2016]. Dostupné z: http://www.eviews.com/illustrated/EViews%20Illustrated.pdf

[13] Brigo, D., Dalessandro, A., Neugebauer, M., Triki, F. *A Stochastic Processes Toolkit for Risk Management* [online]. [citováno 1. 5. 2016]. Dostupné z: https://arxiv.org/pdf/0812.4210.pdf

[14] University of California (USA) and Max Planck Institute for Demographic Research (Germany) *Human Mortality Database* [online]. [data stažena 20. 3. 2016]. Dostupné z: http://www.mortality.org

[15] University of California (USA) and Max Planck Institute for Demographic Research (Germany) *Methods Protocol for the Human Mortality Database* [online]. [citováno 10. 3. 2016]. Dostupné z: http://www.mortality.org/Public/Docs/MethodsProtocol.pdf

[16] Tsay, R. S. *All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models* [online]. [citováno 5. 5. 2016]. Dostupné z: https://cran.r-project.org/web/packages/MTS/MTS.pdf

Seznam obrázků

4.1 Specifická míra úmrtnosti v letech 1970, 1990 a 2010, ČR 32
4.2 Specifická míra úmrtnosti ve věku 50, 60, 70 a 80 let, ČR 32
4.3 Rezidua modelů úmrtnosti pro věk 60 a 80 let, ČR 34
4.4 Specifická míra úmrtnosti v letech 1970, 1990 a 2010, Slovensko . 36
4.5 Specifická míra úmrtnosti ve věku 50, 60, 70 a 80 let, Slovensko . 36
4.6 Specifická míra úmrtnosti ve věku 60 a 70 let 36
4.7 Rezidua modelů úmrtnosti pro věk 60 a 80 let, Slovensko 37
4.8 Parametry LC modelu pro českou a slovenskou populaci 40
4.9 Parametry RH modelu pro českou a slovenskou populaci 42
4.10 Parametry CBD modelu pro českou a slovenskou populaci 44
4.11 Projekce efektů kalendářních roků LC modelu 47
4.12 Projekce specifických efektů v rozšířeném modelu společných faktorů 48
4.13 Projekce specifické míry úmrtnosti LC 48
4.14 Projekce efektu kalendářního roku v RH modelu 51
4.15 Projekce kohortního efektu v RH modelu 52
4.16 Projekce specifické míry úmrtnosti RH 53

Seznam tabulek

2.1 Přehled nejpoužívanějších stochastických modelů 12

4.1 Odhad parametrů metodou maximální věrohodnosti – ČR 33
4.2 Odhad parametrů metodou maximální věrohodnosti – Slovensko . 38
4.3 Odhad parametrů zobecněných LC modelů metodou maximální věrohodnosti 41
4.4 Odhad parametrů zobecněných RH modelů metodou maximální věrohodnosti 43
4.5 Odhad parametrů zobecněných CBD modelů metodou maximální věrohodnosti 44
4.6 p-hodnoty testů reziduálních složek modelů časových řad v modelu LC 47
4.7 Dvacetileté pravděpodobnosti přežití v LC modelu 49
4.8 p-hodnoty testů reziduálních složek modelů $k_1^{[i]}(t)$ v RH modelu . 49
4.9 p-hodnoty testů reziduálních složek kohortních efektů RH modelu 52
4.10 Dvacetileté pravděpodobnosti přežití v RH modelu 53