

NYILATKOZAT

Név: Mózes Árpád Benedek

ELTE Természettudományi Kar, szak: matematika


NEPTUN azonosító: AXQUAH

Szakedolgozat címe:

Általánosított korcsoport–időszak–kohorsz modellek kapocsfüggvényeinek vizsgálata

A **szakedolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2021. május 25.


a hallgató aláírása

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

TERMÉSZETTUDOMÁNYI KAR

**ÁLTALÁNOSÍTOTT KORCSOPORT–IDŐSZAK–KOHORSZ
MODELLEK KAPOCSFÜGGVÉNYEINEK VIZSGÁLATA**

Mózes Árpád Benedek

Matematika BSc, Alkalmazott matematikus szakirány

szakdolgozat

Témavezető:

Próhle Tamás

Valószínűségelméleti és Statisztikai Tanszék



Budapest, 2021.

Tartalomjegyzék

1. Bevezetés	7
2. Alapfogalmak és jelölések	9
3. A modellezéshez kapcsolódó fogalmak	13
3.1. A Gompertz modell	13
4. A GAPC modelles család	16
4.1. Általános modellstruktúra	16
4.2. A vizsgált modellek	18
4.2.1. A Lee–Carter (LC) modell	18
4.2.2. A Renshaw–Haberman (RH) modell	18
4.2.3. A korcsoport–időszak–kohorsz (APC) modell	19
4.2.4. A Cairns–Blake–Dowd (CBD) modellek	20
4.2.5. A Plat modell	21
4.2.6. További modellekről	22
4.2.7. Áttekintés	22
5. Összehasonlítás	23
5.1. Adatok	23

5.1.1. A magyar adatok	24
5.1.2. A lengyel adatok	24
5.2. A StMoMo programcsomag	25
5.2.1. A programcsomag kínálta lehetőségek	26
5.2.2. A programcsomag bővítése és használata	26
5.3. Illeszkedésvizsgálat	30
5.3.1. Futtatás a Currie által javasolt intervallumon	31
5.3.2. Futtatás saját intervallumon	31
5.4. Előrejelzések	32
6. Összefoglalás	33
7. Függelék	34

Köszönetnyilvánítás

Ezúton is szeretném megköszönni mentoromnak, Horváth Gyulának, hogy már korán megismertette velem az élettartam-modellezés alapjait, és ezzel felkeltette érdeklődésemet az aktuárius szakma, és annak tudományos háttere iránt. Neki köszönhető, hogy szakdolgozatom témájával az első perctől szívesen foglalkoztam, és foglalkozom továbbra is.

Köszönöm Prőhle Tamás tanár úrnak a téma ajánlását, a sok segítséget és a hasznos tanácsokat. Köszönettel tartozom szüleimnek fáradhatatlan támogatásukért, valamint mindenki másnak, aki bármilyen formában hozzájárult szakdolgozatom elkészültéhez.

1. fejezet

Bevezetés

Életbiztosítások és járadéktermékek árazásához elengedhetetlen a halandóság jövőbeli alakulásának, a várható élettartamoknak valamilyen becslése. Az utóbbi növekedése több évtizedes trend a fejlett országokban, ami pozitív jelenség, de egyrészt csak véges ideig fenntartható, másrészt kockázatot rejt az előrejelzések készítésében és tartalékolásban az élet- és nyugdíjbiztosítók számára.

Számos, az előrejelzések készítéséhez jól, vagy kevésbé jól használható sztochasztikus halandósági modell létezik. A nagy múltra visszatekintő modellek mellett az elmúlt két évtizedben különböző szerzők több ígéretes modellt is javasoltak. Az ezek közötti hasonlóságot megragadva foglalja őket keretbe az általánosított lineáris modellek (GLM) (példaként lásd: [5, 10]), valamint az általánosított korcsoport-időszak-kohorsz modellek (GAPC) [21] elmélete.

Currie [5] 2016-os cikkében az általánosított lineáris és nemlineáris modellek rendszerezése mellett a megszokottól eltérő kapocsfüggvények kipróbálását javasolja és elemzi hat ország halandósági adatain. Ezzel nem a különböző modelleket hasonlítja egymáshoz, hanem a modellstruktúra egy komponensének megváltoztatásával törekszik a rendelkezésre álló adatokhoz való pontosabb illeszkedésre. Villegas és szerzőtársai [21] még általánosabb modellkeretet írnak le és egyúttal egy könnyen használható programkönyvtárat is bemutatnak, azonban a Currie által javasolt (nem kanonikus) kapocsfüggvények alkalmazásának lehetősége nélkül. Mindez arra sarkall bennünket, hogy a javasolt változatokat a magyar (és egyúttal a lengyel) adatokon is kipróbáljuk, és ezt a Villegas és szerzőtársai által megalkotott kereteken belül tegyük. Az új kapocsfüggvények hazai adatokon való tesztelése mellett tehát a programcsomag szükséges kiterjesztéseinek implementálására is vállalkozunk.

Szakdolgozatunkban először bemutatjuk a szükséges, halandóságot leíró mérőszámokat és fogalmakat, majd egy egyszerű példán keresztül szemléltetjük egy modell felépítését, és a kapcsolódó függvények szerepét. Ezután bemutatjuk az általánosított korcsoport–időszak–kohorsz modelles családot és annak az általunk elemzésre kiválasztott tagjait. Végül bemutatjuk a szóban forgó programcsomagot, megemlítve az általunk eszközölt bővítéseket és elemezzük a kapott eredményeket.

2. fejezet

Alapfogalmak és jelölések

A halandóság leírásának alapvető eszköze a *halandósági ráta*, amely egy adott időszakra és populációra nézve a bekövetkezett halálozások és egy populációrész létszámának arányát jelenti:

$$m = \frac{d}{E},$$

ahol m a halandósági ráta, $d \in \mathbb{N}$ az adott időszakban elhunytak száma, E pedig a populáció létszáma, azaz a kitettség (a elhalálozás rizikójának kitettek).

Ez utóbbit kétféleképpen is definiáljuk, úgy mint

- a vizsgált időszak kezdetén élő egyének száma, ez a *kezdeti kitettség*, (angolul *initial exposure*), melynek jelölése E^0 , valamint
- a vizsgált időszak alatt élő egyének átlagos száma, ez a *központi kitettség*, (angolul *central exposure*), melynek jelölése E^c .

A kettő közötti kapcsolat, ha a vizsgált időszakban elhunyt egyének átlagosan $A > 0$ ideig éltek tehát $E^c = E^0 - (1 - A)d$. Később az $A = \frac{1}{2}$ egyszerűsítő feltevessel élünk, ennek következményeit a fejezet végén taglaljuk.

A vizsgált populáció alapján a halandósági ráták típusait az alábbi szempontok szerint különböztethetjük meg:

- Lakóhely szerint megemlíthetünk országos, regionális, megyei stb. halandósági rátákat is, azonban dolgozatunkban kizárólag országonként elemezzük a rátákat. Currie [5] nyomán szeretnénk egyrészt hat országot: az Amerikai Egyesült Államok, az Egyesült Királyság, Japán, Ausztrália, Svédország és Franciaország

modellillesztési eredményeit (egy könnyebben kezelhető programcsomaggal) reprodukálni, másrészt szeretnénk kiterjeszteni a vizsgálatot Magyarországra és Lengyelországra.

- Nemek szerint megkülönböztetünk uniszex, férfi és női halandósági rátákat. Currie [5] csak férfi adatok alapján végzi az összehasonlítást. Magyarország és Lengyelország esetében ezt a teljes népesség adatain, vagyis uniszex létszámokra és halálesi gyakoriságokra is elvégezzük. A továbbiakban, főleg a programkódokat illetően, rendre a T, M, F jelöléseket használjuk (ang. *total*, *male*, *female*).
- Életkor szerint megkülönböztetünk életkorfüggő vagy életkortól független (nyers) halandósági rátákat. Mint azt a bemutatott modellekből látni fogjuk, vizsgálatunkban fontos szerepet játszanak az életkorfüggő halandósági ráták, melyeket korévekre bontva vizsgálunk. A továbbiakban x jelöli a korcsoporttól való függést, n_a pedig az adott korcsoportok számát.
- A halálozási adatokat naptári évekre lebontva kezeljük. t jelöli a naptári évtől való függést, illetve az adott egy év hosszúságú időszakot, n_y a vizsgált időintervallum éveinek száma.
- Megemlítjük, hogy biztosítási szempontból további változók alapján is felosztható a populáció, így érdekes lehet például a nyugdíjasok, házasok, dohányzók stb. halálozási rátája, ezek vizsgálatára dolgozatunk már nem terjed ki.

A fentiek alapján így írjuk fel a kezdeti és központi halandósági rátákat:

$$m_{x,t}^0 = \frac{d_{x,t}}{E_{x,t}^0}, \quad m_{x,t}^c = \frac{d_{x,t}}{E_{x,t}^c}, \quad x \in \{1, \dots, n_a\}, \quad t \in \{1, \dots, n_y\},$$

ahol $m_{x,t}$ a t . naptári évben x éves egyénekre vonatkozó (kezdeti vagy központi) halandósági ráta, $d_{x,t}$ a t . naptári év során x évesen elhunyt egyének száma. $E_{x,t}^0$ a t . naptári év kezdetén x évesek száma, $E_{x,t}^c$ pedig a t . naptári évben x . életévét betöltött átlagos népesség száma.

Ezzel hallgatólagosan feltettük, hogy az x évesen elhunyt egyén pont az év kezdetén töltötte be az x . életévét, ami kevés kivételtől eltekintve nyilván nem igaz. A gyakorlatban egyrészt a számlálóban szerepelnek olyan elhunytak, akik a naptári év kezdetén még csak $x - 1$ évesek voltak, másrészt a nevezőben szerepelnek olyan egyének, akik bár a t . évben hunytak el, de ekkor már $x + 1$ évesen, s emiatt nem szerepelnek a $d_{x,t}$ számlálóban. Ennek kiegyenlítésére a KSH az ún. Böckh-formulát [17] alkalmazza. Az ebből eredő különbség nem számottevő. [22]

Halálozási valószínűség alatt annak a valószínűségét értjük, hogy egy x . életévét éppen betöltött egyén a következő egy év során meghal, azaz

$$q(x) = \mathbb{P}(L < x + 1 \mid L \geq x) \quad (x \in \mathbb{N}),$$

ahol L az egyén élettartamát kifejező nemnegatív valószínűségi változó.

Ahogy az előző bekezdésben leírtuk, feltételezzük, hogy az x évesen elhunyt egyén pont az év kezdetén töltötte be az x . életévét; tehát $q_{x,t}$ -t úgy értelmezzük, mint a t . naptári év kezdetével az x . életévét betöltő egyén halálozási valószínűsége (a t . év során).

Halálozási intenzitás (vagy házárdráta, angolul *force or mortality* vagy *hazard rate*) alatt egyfajta pillanatnyi halálozási valószínűséget értünk. Ez azt jelenti, hogy egy rövid $(t, t + dt)$ időintervallumban egy μ halálozási intenzitásnak kitett egyén halálozási valószínűségét μdt jól közelíti (kis dt esetén). Precízebben [12, 22]:

$$\mu(y) = \lim_{\varepsilon \rightarrow 0+} \frac{\mathbb{P}(L < y + \varepsilon \mid L \geq y)}{\varepsilon} \quad (y \geq 0).$$

Feltehető, hogy a halálozási intenzitás – mint időpontbeli mennyiség – a népesség tekintetében lassan és folytonosan differenciálható módon változik, mind időben, mind életkor szerint. Egy általános megközelítés [3, 5], hogy konstansnak tekintjük naptári éven belül, és koréven belül. Ennél fogva $\mu_{x,t}$ -vel jelöljük, és úgy értelmezzük, mint a t . évben az x életkorra vonatkozó konstans halálozási intenzitás. Ez a diszkrét (évenkénti, illetve korévenkénti bontásban rendelkezésünkre álló) adatokra nézve a túlélési függvény¹ exponenciális interpolációját jelenti.

Ennek további előnye, hogy következik belőle az alábbi összefüggés:

$$q_{x,t} = 1 - \exp(-\mu_{x,t}) \quad \text{és} \quad \mu_{x,t} = -\log(1 - q_{x,t}).$$

Ha $d_{x,t}$ -t valószínűségi változóként kezeljük, annak eloszlása rögzített kezdeti kitettség mellett binomiális eloszlás:

$$D_{x,t} \sim \text{Bin}(E_{x,t}^0, q_{x,t}),$$

rögzített központi kitettség mellett pedig Poisson-eloszlás:

$$D_{x,t} \sim \text{Poisson}(E_{x,t}^c \mu_{x,t}).$$

¹ $G(y) = \mathbb{P}(L \geq y) \quad (y \geq 0)$

Ebből az alábbi log-likelihood függvény adódik a Possion-eloszlás esetében:

$$\ell(q) = \sum_{x=1}^{n_a} \sum_{t=1}^{n_y} \left(\log \left(\frac{E_{x,t}^0}{d_{x,t}} \right) + d_{x,t} \log q_{x,t} + (E_{x,t}^0 - d_{x,t}) \log(1 - q_{x,t}) \right),$$

és az alábbi a binomiális eloszlás esetében:

$$\ell(\mu) = \sum_{x=1}^{n_a} \sum_{t=1}^{n_y} \left(-E_{x,t}^c \mu_{x,t} + d_{x,t} (\log E_{x,t}^c + \log \mu_{x,t}) - \log(d_{x,t}!) \right),$$

melyekből elemi differenciálszámítással a maximum likelihood becsléseik:

$$\hat{q}_{x,t} = \frac{d_{x,t}}{E_{x,t}^0} = m_{x,t}^0 \quad \text{és} \quad \hat{\mu}_{x,t} = \frac{d_{x,t}}{E_{x,t}^c} = m_{x,t}^c.$$

Az $E^c = E^0 - (1 - A)d$ egyenletben az $A = \frac{1}{2}$ feltevés, vagyis az $E_{x,t}^0 = E_{x,t}^c + \frac{1}{2}d_{x,t}$ feltevés eltér az említett exponenciális interpoláció feltételezésétől, azonban belátható, hogy az eltérés elhanyagolhatóan csekély.

Az előző, kitettségekre vonatkozó egyenletből az alábbi összefüggés következik:

$$m_{x,t}^0 = \frac{m_{x,t}^c}{1 + \frac{1}{2}m_{x,t}^c},$$

az exponenciális interpolációból és a log-likelihood függvényekből viszont az alábbi összefüggés *következne*:

$$m_{x,t}^0 = 1 - \exp(-m_{x,t}^c).$$

A nulla körüli Taylor-soraikból azonban megállapítható, hogy

$$\frac{m_{x,t}^c}{1 + \frac{1}{2}m_{x,t}^c} - (1 - \exp(-m_{x,t}^c)) \approx \frac{1}{12}(m_{x,t}^c)^3,$$

azaz $A = \frac{1}{2}$ feltevés mellett

$$m_{x,t}^0 \approx 1 - \exp(-m_{x,t}^c)$$

és

$$\hat{q}_{x,t} \approx 1 - \exp(-\hat{\mu}_{x,t})$$

jó közelítés, így a továbbiakban ezt alkalmazzuk.

A fejezet definícióit illetően Vékás Péter [22, 3. fejezet] magyar nyelvű módszertani összefoglalójának főbb pontjait követtük, kiegészítve a dolgozatunkra vonatkozó megállapításokkal és jelölésekkel.

3. fejezet

A modellezéshez kapcsolódó fogalmak

3.1. A Gompertz modell

Kapocsfüggvények és modellstruktúra ismertetése egy egyszerű lineáris modell példáján keresztül

Benjamin Gompertz brit matematikus és aktuárius már 1825-ben [7] megfigyelte, hogy a felnőttkori halálozási intenzitás logaritmikus skálán megközelítőleg lineáris. A 3.1. ábrán látható példa szerint ez a magyar halálozási adatokra is fennáll (az 1960. és 2017. évben). A Gompertz-féle halandósági törvény (ang. *mortality law*) a következő formában is ismeretes [19]:

$$\mu_{x,t} = e^{\theta_0 + \theta_1 x}$$

Amiből Poisson-eloszlást feltételezve

$$\eta_{x,t} = \log \mathbb{E}[D_{x,t}] = \log E_{x,t}^c + \log \mu_{x,t} = \log E_{x,t}^c + \theta_0 + \theta_1 x$$

felírásával már egy lineáris modellt definiáltunk $D_{x,t}$ valószínűségi változóra Poisson-eloszlással, logaritmikus kapocsfüggvénnyel, és $\eta_{x,t}$ szisztematikus komponenssel (becslőfüggvénnyel). [5]

A Poisson-eloszlás esetében a logaritmusfüggvény a kanonikus kapocs a valószínűségi változó és a szisztematikus komponens között. A következőkben végigvesszük, hogy Currie [5] cikkében mi alapján, és hogyan javasol különböző kapocsfüggvényeket.

A Poisson-esetben a logaritmusfüggvény mellett szól

- az adatok mintázata (legalábbis magasabb életkorokra), mint azt a 3.1. ábra szemlélteti,
- a logaritmusfüggvény azon tulajdonsága, hogy a pozitív félegyenesről a teljes számegyenesre képez,
- általa egyszerűbb a maximum likelihood egyenletek megoldása.

Ugyanakkor javasolja a logit függvényt:

$$\text{logit}(q) = \log\left(\frac{q}{1-q}\right), \quad 0 < q < 1.$$

A logit kapocsfüggvény használata azért nem magától értetődő, mert értelmezési tartománya a szűkebb $(0, 1)$ intervallum, azonban megállapítható, hogy a halálozási intenzitás még a legmagasabb életkorokban is ezen az intervallumon belül van.

Visszatérve a Gompertz-féle halandósági törvényhez, azt a

$$q_{x,t} \approx 1 - \exp(-\mu_{x,t})$$

közelítésbe helyettesítve $\log(-\log(1 - q_{x,t})) \approx \theta_0 + \theta_1 x$ adódik.

Az bal oldalon szereplő

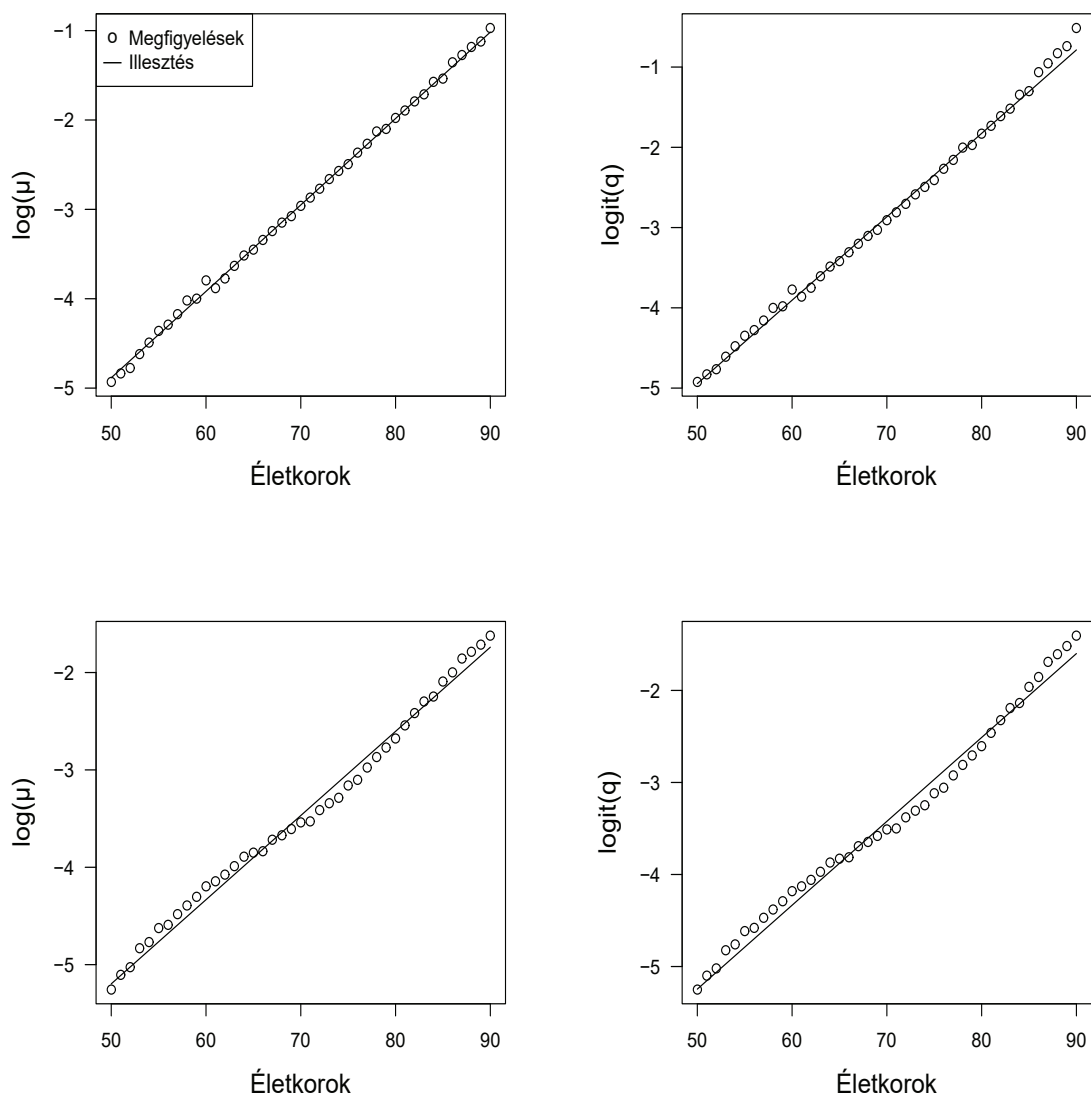
$$\text{cloglog}(q) = \log(-\log(1 - q)), \quad 0 < q < 1$$

függvényt magyarul komplementis log-log függvénynek nevezzük (angolul *complementary log-log function*). Binomiális eloszlás feltételezésével ekkor a $D_{x,t}$ -ből származtatott $Q_{x,t} = D_{x,t}/E_{x,t}^0$ valószínűségi változóra

$$\eta_{x,t} = \text{cloglog}\mathbb{E}[Q_{x,t}] = \text{cloglog}q_{x,t} = \theta_0 + \theta_1 x.$$

Ezzel ismét egy, az általánosított lineáris modellek [3] terminológiájába illő modellt kapunk, $D_{x,t}/E_{x,t}^0$ valószínűségi változóra binomiális eloszlással, komplementis log-log kapocsfüggvénnyel, és $\eta_{x,t} = \theta_0 + \theta_1 x$ szisztematikus komponenssel.

Ennek ellenére a binomiális eloszlás esetén a kanonikus kapocsfüggvény a logit függvény. Ennek okai megegyeznek a Poisson-eloszlás esetén felsorolt, a logaritmikus kapocsfüggvény mellett szóló érvekkel.



3.1. ábra. Megfigyelt halálozási intenzitások és a Gompertz modellel illesztett egyenese-
sek. A bal oldalon Poisson-eloszlást és logaritmikus kapocsfüggvényt, a jobb oldalon
binomiális eloszlást és logit kapocsfüggvényt feltételezve. Magyar adatok; a felső
sorban az 1960. év férfi népessége, az alsó sorban a 2017. év teljes népessége.

4. fejezet

A GAPC modellesalád

Ebben a fejezetben az általánosított korcsport–időszak–kohorsz (GAPC) modellesalád elméletét ismertetjük, melyet Villegas és szerzőtársai [21] javasoltak. Az általános esetben az általuk leírt felépítést követjük, majd bemutatjuk a modellesalád összehasonlításunkban szereplő tagjait.

4.1. Általános modellstruktúra

A modell alkalmazásához feltesszük, hogy a vizsgált időszakra és korévekre ismert a halálozások $d_{x,t}$ száma, mint a $D_{x,t}$ valószínűségi változó megfigyelt értékei, valamint a kezdeti ($E_{x,t}^0$) vagy központi ($E_{x,t}^c$) kitettségek

- minden egyes korévre: $x \in \{1, 2, \dots, n_a\}$,
- és minden egyes naptári évre: $t \in \{1, 2, \dots, n_y\}$.

Megfigyeléseink tehát $n_a \times n_y$ dimenziójú mátrixot alkotnak. Amennyiben a kitettségnek csak egyik változata ismert, a 2. fejezetben részletezett $E_{x,t}^0 = E_{x,t}^c + \frac{1}{2}d_{x,t}$ feltevés alapján számítjuk ki a másikat. Jelölje $n_c = n_a + n_y - 1$ a kohorszok számát.

Egy általánosított korcsport–időszak–kohorsz sztochasztikus halandósági modell ekkor az alábbi négy részből áll.

1. A halálozások számának feltételezett eloszlása (vagy valószínűségi komponens), amely

$$D_{x,t} \sim \text{Poisson}(E_{x,t}^c \mu_{x,t})$$

vagy $D_{x,t} \sim \text{Bin}(E_{x,t}^0, q_{x,t})$,

ami által rendre $\mathbb{E}(D_{x,t}/E_{x,t}^c) = \mu_{x,t}$ vagy $\mathbb{E}(D_{x,t}/E_{x,t}^0) = q_{x,t}$.

2. A szisztematikus komponens, amely a három névadó hatás valamely függvényeként jellemzi a halandóságot. Általános alakja:

$$\eta_{x,t} = \alpha_x + \sum_{i=1}^N \beta_x^{(i)} \kappa_t^{(i)} + \beta_x^{(0)} \gamma_{t-x}.$$

Ahol:

- α_x időben változatlan *korcsoport-hatás*, amely ezáltal a halandóság általános alakulását ragadja meg az adott korévekre nézve,
- $N \geq 0$ egész szám, amely a halandóság időbeli alakulását idősorok formájában leíró $\kappa_t^{(i)}$ ($i = 1, \dots, N$) *mortalitási indexek* számát adja meg,
- γ_{t-x} *kohorszhatás*, amely a $t - x$ kohorszra jellemző, a halandóságot befolyásoló hatásokat reprezentálja,
- $\beta_x^{(i)}$ az $i = 1, \dots, N$ esetben az i -edik mortalitási indexre, $i = 0$ esetben a kohorszhatásra vonatkozó *életkorfüggő érzékenységi együttható*. Ez fejezi ki, hogy az említett hatások különböző korcsoportok halandóságát milyen mértékben befolyásolják.

3. A kapocsfüggvény, amely kapcsolatot teremt az előző két komponens között:

$$g\left(\mathbb{E}\left(\frac{D_{x,t}}{E_{x,t}}\right)\right) = \eta_{x,t}.$$

Itt $E_{x,t}$ alatt Poisson-eloszlás esetén a központi, binomiális eloszlás esetén a kezdeti kitettséget értjük. Dolgozatunk fő tárgya az alábbi négy kapocsfüggvény-eloszlás kombináció összehasonlítása:

- e -alapú logaritmusrfüggvény Poisson-eloszlás mellett (kanonikus),
- logit függvény Poisson-eloszlás mellett,
- logit függvény binomiális eloszlás mellett (kanonikus),
- cloglog függvény binomiális eloszlás mellett.

Ezeket a függvényeket részletesebben az előző fejezetben mutattuk be.

4. Az egyértelműsítő (identifikációs) megkötések. A legtöbb modell csak egy bizonyos, a paramétereire alkalmazott transzformáció erejéig egyértelmű, ezért a paraméterek becsléséhez megkötéseket kell tennünk. Ezt a transzformációt az első vizsgált modell esetében (Lee–Carter) szemléltetjük, a továbbiakban csak az általunk is használt megkötéseket tüntetjük fel.

4.2. A vizsgált modellek

Ebben a szakaszban bemutatjuk azokat a modelleket melyekre nézve a következő fejezetben a különböző kapocsfüggvényeket összehasonlítjuk. Az általános modellstruktúrához hasonlóan feltüntetjük egyrészt a modellek szisztematikus komponensét, ezzel mutatva, hogy hogyan illeszkednek a GAPC model családba, másrészt megadjuk a szükséges egyértelműsítő megkötéseket.

4.2.1. A Lee–Carter (LC) modell

A Ronald D. Lee és Lawrence R. Carter által 1992-ben [13] javasolt, mára klasszikussá vált modell Brouhns és szerzőtársai [1] által bevezetett, hibatagok nélküli változata a következő:

$$\eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)}.$$

A hibatagra ebben az esetben azért nincs szükség, mert a becsült értékek szórását a feltételezett eloszlás adja meg.

Könnyen belátható, hogy a fenti egyenlet invariáns a paraméterek alábbi transzformációjára:

$$\{\alpha, \beta, \kappa\} \mapsto \left\{ \alpha + c_1 \beta, \frac{1}{c_2} \beta, c_2 (\kappa - c_1 \mathbf{1}) \right\} \quad c_1, c_2 \in \mathbb{R}, c_2 \neq 0.$$

Lee és Carter az egyértelműség érdekében a következő megkötéseket javasolják:

$$\sum_{x=1}^{n_a} \beta_x^{(1)} = 1, \quad \sum_{t=1}^{n_y} \kappa_t^{(1)} = 0.$$

Ez a modell nem tartalmaz kohorszhatást, a halandóság alakulását egyetlen életkorfüggő érzékenységgel módosított mortalitási index jellemzi. Ez a szorzat az oka annak, hogy a Lee–Carter nem lineáris modell. Brouhns és szerzőtársai [1] eredetileg Poisson-eloszlást és logaritmikus kapocsfüggvényt feltételeznek, központi kitettségekkel a halálozási intenzitást célozzák.

4.2.2. A Renshaw–Haberman (RH) modell

Renshaw és Haberman [18] a Lee–Carter modell kohorszhatással bővített változatát javasolják:

$$\eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \beta_x^{(0)} \gamma_{t-x}.$$

A fenti, a kohorszhatást is életkorfüggő érzékenységgel jellemző modell numerikusan instabil, ezért erről lemondva a szerzők később [8] az alábbi megkötéssel egyszerűsített változatot javasolják:

$$\beta_x^{(0)} = 1, \quad \eta_{x,t} = \alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \gamma_{t-x}.$$

Dolgozatunkban is az utóbbi változatot használjuk. Hunt és Villegas [10] rámutattak, hogy az egyszerűsített változatban is léphetnek fel problémák a paraméterbecslés során. Ennek megoldására a Newton–Raphson-módszert javasolják a log-likelihood függvény maximalizálására, ahogyan azt Brouhns és szerzőtársai [1] tették a Lee–Carter modell esetében.

Az egyszerűsített modellváltozatban négy helyett csak az alábbi három egyértelműsítő megkötés szükséges:

$$\sum_{x=1}^{n_a} \beta_x^{(1)} = 1, \quad \sum_{t=1}^{n_y} \kappa_t^{(1)} = 0, \quad \sum_{c=1}^{n_c} \gamma_i = 0.$$

A szerzők eredetileg ebben az esetben is Poisson-eloszlást és logaritmikus kapcsolási függvényt, illetve központi kitettségeket feltételeznek a halálozási intenzitást célozva.

4.2.3. A korcsoport–időszak–kohorsz (APC) modell

A korcsoport–időszak–kohorsz (angolul *Age–Period–Cohort*, röviden APC) az orvostudomány és a demográfia területén tekint vissza nagyobb múltra, az aktuárius szakirodalomban Currie (2006) [4] cikke után terjedt el. [21] Tekintható a Renshaw–Haberman modell alváltozatának $\beta_x^{(1)} = 1$ és $\beta_x^{(0)} = 1$ feltételek mellett:

$$\eta_{x,t} = \alpha_x + \kappa_t^{(1)} + \gamma_{t-x}.$$

A három hatást tehát egymástól függetlenül, életkorfüggő érzékenységek nélkül kezeli. Leggyakoribb egyértelműsítő megkötései:

$$\sum_{t=1}^{n_y} \kappa_t^{(1)} = 0, \quad \sum_{c=1}^{n_c} \gamma_i = 0, \quad \sum_{c=1}^{n_c} c\gamma_i = 0.$$

Struktúrájából adódóan (a továbbiakhoz hasonlóan) lineáris modell. Currie (2006) [4] Poisson-eloszlást és logaritmikus kapcsolási függvényt feltételezve mutatja be, becslése központi kitettségeket alkalmazva a halálozási intenzitást célozza.

4.2.4. A Cairns–Blake–Dowd (CBD) modellek

Cairns és szerőtársai [2] cikkükben a következő, két mortalitási indexet tartalmazó modellt javasolják az időskori halandóság¹ előrejelzésére:

$$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)},$$

ahol \bar{x} az előforduló korcsoport-indexek átlaga, valamint $\beta_x^{(1)} = 1$, $\beta_x^{(2)} = x - \bar{x}$ előre meghatározott érzékenységi együtthatók.

A CBD modellek paramétereinek becslése eredetileg a halálozási valószínűségeket célozza logit kapocsfüggvénnyel. Találunk példát Poisson-eloszlás feltételezésére központi kitettségekkel [3], illetve binomiális eloszlás feltételezésére kezdeti kitettségekkel [8].

Ebben az alapmodellben nem szerepel életkorhatás és kohorszhatás, és egyértelműsítő megkötésekre sincs szükség. Cairns és szerőtársai [3] a modell több kiterjesztését is javasolják, ebből (az eredeti mellett) az alábbi kettőt használjuk majd.

M6: CBD modell kohorszhatással

A fent hivatkozott [3] cikk alapján M6-tal (vagy CBD(C) módon) jelölt modell az eredetinél egy additív kohorszhatással több. Szisztematikus komponense:

$$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x},$$

ahol

$$\beta_x^{(1)} = 1, \quad \beta_x^{(2)} = x - \bar{x}, \quad \beta_x^{(0)} = 1.$$

Ezzel a bővítéssel két egyértelműsítő megkötést is kell tennünk:

$$\sum_{c=1}^{n_c} \gamma_c = 0, \quad \sum_{c=1}^{n_c} c \gamma_c = 0.$$

M7: négyzetes CBD modell kohorszhatással

Az M7-tel (vagy CBD(QC) módon) jelölt változatban a szerzők a kohorszhatás mellett, egy harmadik mortalitási indexszel, négyzetes életkor-hatás bevezetését javasolják:

¹60 évesnél magasabb életkorok

$$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + [(x - \bar{x})^2 - \hat{\sigma}_x^2] \kappa_t^{(3)} + \gamma_{t-x},$$

ahol $\hat{\sigma}_x^2$ az $(x - \bar{x})^2$ értékek átlaga,

$$\beta_x^{(1)} = 1, \quad \beta_x^{(2)} = (x - \bar{x}), \quad \beta_x^{(3)} = [(x - \bar{x})^2 - \hat{\sigma}_x^2], \quad \beta_x^{(0)} = 1.$$

Ezzel együtt a szükséges megkötések száma háromra nő:

$$\sum_{c=1}^{n_c} \gamma_c = 0, \quad \sum_{c=1}^{n_c} c \gamma_c = 0, \quad \sum_{c=1}^{n_c} c^2 \gamma_c = 0.$$

4.2.5. A Plat modell

Dolgozatunkban az összehasonlítást elvégezzük a Currie-nél [5] nem szereplő, Plat [15] által javasolt modellen is, melyben a CBD modellt kohorszhatással és – a Lee–Carter modellhez hasonlóan – additív korcsoport-hatással ötvözi:

$$\eta_{x,t} = \alpha_x + \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + (\bar{x} - x)^+ \kappa_t^{(3)} + \gamma_{t-x},$$

ahol $(\bar{x} - x)^+ = \max \{0, \bar{x} - x\}$,

$$\beta_x^{(1)} = 1, \quad \beta_x^{(2)} = (x - \bar{x}), \quad \beta_x^{(3)} = (\bar{x} - x)^+, \quad \beta_x^{(0)} = 1.$$

Időskori halandóság vizsgálata esetén a szerző a harmadik mortalitási index elhagyását javasolja. Mivel az összehasonlítást ilyen életkorokon végezzük majd, ezért mi is így teszünk. Ezzel a szisztematikus komponens az alábbi módon egyszerűsödik:

$$\eta_{x,t} = \alpha_x + \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x}.$$

Ez az egyszerűsített változat tulajdonképp az M6 modell korcsoport-hatással való bővítése. Egyértelműsítő megkötései:

$$\sum_{t=1}^{n_y} \kappa_t^{(1)} = 0, \quad \sum_{t=1}^{n_y} \kappa_t^{(2)} = 0, \quad \sum_{c=1}^{n_c} \gamma_c = 0, \quad \sum_{c=1}^{n_c} c \gamma_c = 0.$$

Plat eredetileg Poisson-eloszlást és logaritmikus kapcsolási függvényt feltételez, központi kitettségekkel a halálozási intenzitást célozza.

4.2.6. További modellekről

Az alábbi két modellt nem tárgyaljuk dolgozatunkban.

M4: 2-d P-spline modell

A Cairns és szerzőtársai [3] cikkében M4 jelöléssel szereplő 2-dimenziós P-spline modell felépítése alapján nem tartozik a GAPC modelleszaládba, bár nagyban összefügg az általánosított lineáris modellek elméletével. Illesztése egy külön elemzést igényelne, emellett kritikus pontja, hogy retrospektív értelemben jól működik, de előrejelzési célokra való alkalmazhatósága kérdéses.

M8: CBD modell korcsoport-függő kohorszhatással

A CBD modellnek egy Currie-nél [5] is szereplő változata az M8 (vagy CBD(C_δ)) jelölésű, korcsoport-függő additív kohorszhatást tartalmazó változat:

$$\eta_{x,t} = \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + (\delta - x) \gamma_{t-x},$$

ahol δ egy szintén becsült, konstans paraméter, a javasolt megkötés $\sum_{c=1}^{n_c} \gamma_c = 0$.

A δ paraméter becsléséből adódó nehézségek miatt nem tűnik használhatónak előrejelzések készítéséhez. [5, 371. o.]

4.2.7. Áttekintés

Rövidítés	Szisztematikus komponens	Effektív paraméter ¹
LC	$\alpha_x + \beta_x^{(1)} \kappa_t^{(1)}$	$2n_a + n_y - 2$
RH	$\alpha_x + \beta_x^{(1)} \kappa_t^{(1)} + \gamma_{t-x}$	$2n_a + n_y + n_c - 3$
APC	$\alpha_x + \kappa_t^{(1)} + \gamma_{t-x}$	$n_a + n_y + n_c - 3$
CBD	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)}$	$2n_y$
M6	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x}$	$2n_y + n_c - 2$
M7	$\kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + [(x - \bar{x})^2 - \hat{\sigma}_x^2] \kappa_t^{(3)} + \gamma_{t-x}$	$3n_y + n_c - 3$
Plat	$\alpha_x + \kappa_t^{(1)} + (x - \bar{x}) \kappa_t^{(2)} + \gamma_{t-x}$	$n_a + 2n_y + n_c - 4$

¹Effektív paraméterek száma alatt a modell szisztematikus komponenséből is leolvasható paraméterszámnak és az egyértelműsítő megkötések számának különbségét értjük.

5. fejezet

Összehasonlítás

5.1. Adatok

Az felhasznált adatok forrása a Human Mortality Database (röviden HMD) [9], ahol országonként elérhetőek többek között a halálozási és kitettségi táblázatok, halálozási ráták, népességszámok. Ezek évenkénti korosztályos bontásban is rendelkezésünkre állnak, azzal a megkötéssel, hogy a 109 éves kor feletti értékeket egy közös 110+ jelölésű kategória tömöríti. Az adatok elérhetősége (a legkorábbi és a legutóbbi évek tekintetében) országonként változó. Currie [5] ezért az összehasonlítást az 1960–2009. intervallumon végzi, férfi adatokon. A vizsgált életkorokat – a nyugdíj- és járadék-számítás szempontjából érdekes – 50 éves kortól 90 éves korig terjeszti ki (szemben a Cairns és szerzőtársai (2009) [3] által vizsgált 60–89 intervallummal).

Összehasonlításunkban először maradunk a Currie [5] által javasolt éveknél és kor éveknél, majd a magyar és lengyel adatokra való illeszkedést megvizsgáljuk egy saját intervallumon is: 65–90 életkorokra az 1988–2017. évekre, uniszex adatokon. Ennek okait alább felsoroljuk.

- Az előző fejezetben bemutatott CBD és Plat modellek alkalmazását a szerzők 60 évesnél magasabb életkorokra javasolják, valamint a nyugdíjkorhatár Magyarországon jelenleg 65 év, Lengyelországban férfiak esetében 65, nők esetében 60 év.
- A rendelkezésünkre állnak a 2009-esnél frissebb adatok.
- Egy, például 10 éves előrejelzés készítéséhez nem célszerű egészen 1960-tól kezdenünk az illesztést.

- Uniszex adatokat használunk, mivel az Európai Unió erre vonatkozó irányelve értelmében a járadékbiztosítások díjkalkulációjában tilos a nemek szerinti megkülönböztetés. [22]

5.1.1. A magyar adatok

A HMD magyarországi adatainak forrása a Központi Statisztikai Hivatal, azonban ezeket nem változtatás nélkül tartalmazza, pontosabban nem minden időszakra.

Ahogy az adatbázis dokumentációjában Németh László és szerzőtársai [14] is leírják, jóval régebbi statisztikák is rendelkezésre állnak, azonban ezek minősége, teljessége erősen változó. Ráadásul az ország jelentős területi változásai megnehezíték a halálozási tendenciák értelmezését, ezért a táblázatok az 1950. évvel kezdődnek, azzal a figyelmeztetéssel, hogy az 1950–1959. évi adatok így is óvatossággal kezelendők.

A hivatalos népességi becslések a népszámlálások utáni becslések, és nem voltak újraszámolva későbbi népszámlálások alapján, ráadásul az 1960. és 2001. közötti időszakban nem vették figyelembe a nemzetközi migrációt. Ez történelmi okokból kifolyólag törést okozott a népességszám folytonosságában az 1990. és a 2000. évi népszámlálás alkalmával

Az említett okok miatt az adatbázis az 1960. és 2001. közötti intervallumon a HMD módszertanának [11] segítségével becsült adatokat tartalmaz, így megszakítások nélkül vizsgálhatjuk a tendenciákat. Másrészt pedig azért is fontos ezt megjegyeznünk, mert számításainkban a halálozások számát diszkrét valószínűségi változókkal modellezzük, míg az említett időszak nem csak a kis létszámú korcsoportokra (80 évtől felfelé), hanem minden korévre tört számokat tartalmaz.

Ezt a programkód futtatása után figyelmeztető üzenetek is jelzik, de a bemeneti értékek kerekítésével eredményt így is kapunk. Mivel elsődleges célunk a különböző kapcsolási függvények összehasonlítása, nem a minél pontosabb előrejelzések készítése, ezért az ebből a kerekítésből adódó pontatlanságtól eltekintünk.

5.1.2. A lengyel adatok

Lengyelország adatainak forrása főképp a lengyel Központi Statisztikai Hivatal (lengyelül Główny Urząd Statystyczny, röviden GUS) részben pedig Agnieszka Fihel gyűj-

tése az állami levéltárból.

Ahogy az adatbázis dokumentációja [6] részletesen leírja, a lengyel adatokban is vannak kiigazítva közölt, illetve körültekintéssel kezelendő szakaszok. Bár 1950-ben is volt népszámlálás, a köztes évek becslését történelmi okok mellett tovább nehezítik az 1960. évi népszámlálás problémái. Az első év az adatbázisban 1958.

Az említett 1960. évben az eredeti adatok egy jelentős hullámvölgyet mutatnak a 20–22 éves férfi lakosság körében. Ennek oka ismeretlen, de valószínűsíthető, hogy a férfi sorkatonák maradtak ki a népességszámlálásból. A GUS emiatt átdolgozta az említett népszámlálási év adatait, az adatbázis is ez alapján készült.

A HMD módszertanával [11] kiigazított további időintervallumok, a teljesség igénye nélkül:

- 1960–1969. az 1970-es népszámlálásig regisztrálatlan emigráció miatt,
- 1988–2000. az 1990-es években nagy mértékű regisztrálatlan kivándorlás miatt.

A rendelkezésre álló időszak alatt Lengyelországban a demográfiai statisztikákhoz négy különböző népességdefiníciót használtak, emellett 1958-tól 1994-ig a nemzetközi szabványtól eltérő definíciót alkalmaztak a születésre és csecsemőhalálozásra. Utóbbi miatt a GUS 1970-ig visszamenőleg korrigálta a csecsemőhalandóságra vonatkozó adatokat.

5.2. A StMoMo programcsomag

Az R programcsomag [16] **StMoMo** programkönyvtára [21] az általánosított GAPC modelles család tömörségét a **gnm** programkönyvtár [20] hathatós illesztési függvényével ötvözi, így számos halandóság-előrejelző modell implementálásához nyújt hatékony eszközt. A modelleket a 4. fejezetben már ismertettük ugyanezen keretrendszerben, összehasonlításukhoz a **StMoMo** 0.4.1-es verzióját, illetve annak saját kiegészítéseit használjuk.¹

¹A programcsomag neve az angol *Stochastic Mortality Modeling* akronímája, kiejtése (angolul) *Saint Momo*. Momo latin-amerikai számos részén a karneválok bálkirályának neve. [21]

5.2.1. A programcsomag kínálta lehetőségek

A **StMoMo** programcsomagon kívül, annak megjelenése előtt is léteztek különböző csomagok egy vagy több említett modell implementálásával, azonban ezekből hiányoztak, vagy nem voltak bármely definiált modellre alkalmazhatóak olyan eszközök, mint például az előrejelzések, szimulációk készítése.

A **StMoMo**-val mindezekre lehetőség van. A GAPC keretein belül bármilyen absztrakt modellt egyszerűen definiálhatunk (kivéve a paraméteres és nemparaméteres érzékenységi együttthatók egyidejű használatát), így a GAPC család tetszés szerint bővíthető. A csomag ezáltal korábban nem implementált modelleket is magába foglal.

Vizsgálatunk szempontjából legnagyobb hátránya, hogy a Currie [5] által javasolt kapocsfüggvények közül csak a két kanonikus változatot tartalmazza: a logaritmusfüggvényt Poisson-eloszlás esetén, és a logit függvényt binomiális eloszlás esetén, ezért szakdolgozatunk keretén belül kibővítjük a programkönyvtárat.

5.2.2. A programcsomag bővítése és használata

Ebben a fejezetben bemutatjuk a programcsomag használatát, kiemelve azokat a pontokat, ahol a szükséges bővítéseket létrehozzuk. Szögletes zárójelben a módosított szkriptfájl nevét tüntetjük fel, a teljes listát a függelék tartalmazza.

A **StMoMo**-ban nincs külön függvény külső adatok importálásához, ezért a **demography** programcsomag [23] `hmd.mx` függvényét hívjuk segítségül. Ezzel közvetlenül a Human Mortality Database [9] adatait töltjük le. Ehhez felhasználónév és jelszó, azaz előzetes regisztráció szükséges. A elérhető populációkat (országokat) a HMD weboldalán, illetve a **demography** R-dokumentációjában rögzített módon rövidítések azonosítják.

```
HUdata <- hmd.mx(country = "HUN", username = "...",  
  password = "...", label = "Hungary")
```

Az így nyert adatokon futtatnunk kell a **StMoMoData** függvényt. Ezzel egyrészt áttérünk a programcsomag saját adattípusára (**StMoMoData**), másrészt szétválasztjuk a női, férfi, uniszex adatsorokat.

```
HUFemaleData <- StMoMoData(HUdata, series = "female")  
HUMaleData <- StMoMoData(HUdata, series = "male")  
HUTotalData <- StMoMoData(HUdata, series = "total")
```

A következő lépés a GAPC modellek definiálása. Ehhez általánosan a **StMoMo** függvény használható.

```
StMoMo(link = c("log", "logit", "logitP", "cloglog"),
  staticAgeFun = TRUE, periodAgeFun = "NP", cohortAgeFun
  = NULL, constFun = function(ax, bx, kt, b0x, gc, wxt,
  ages) list(ax = ax, bx = bx, kt = kt, b0x = b0x, gc =
  gc))
```

- ahol a **link** a választott kapocsfüggvényt és a feltételezett eloszlást határozza meg,
- **staticAgeFun** logikai változó, ami azt jelzi, hogy tartalmaz-e a modell α_x időben változatlan korcsoport-hatást,
- **periodAgeFun** egy N hosszú lista, amely a $\beta_x^{(i)}$ mortalitási indexekre vonatkozó életkorfüggő érzékenységi együtthatók definícióit tartalmazza,
- **cohortAgeFun** a $\beta_x^{(0)}$ kohorszhatásra vonatkozó életkorfüggő érzékenységi együttható definíciója,
- **constFun** pedig az identifikációs megkötéseket, mint a paraméterek függvényét tartalmazza.

Programozás szempontjából a függvény eredetileg a **link** = "log" vagy "logit" alapján következtet log-Poisson vagy logit-binomiális modellre. Mint arra korábban utaltunk, a **StMoMo** a **gnm** csomag illesztési függvényét használja. A értelmezést megkönnyítő szöveges formula előállítás mellett a **StMoMo** függvény fő feladata az ehhez szükséges **gnm**-formula, vagyis a **gnm** függvény argumentumának megkonstruálása. Ehhez mi hozzáadjuk a logit-Poisson és cloglog-binomiális modelleknek megfelelő "logitP" és "cloglog" bemeneti lehetőségeket, és a hozzájuk tartozó **gnm**-konstrukciókat. [lásd: StMoMo.R]

Szinte az összes általunk vizsgált, nevezetes modell saját definiáló függvénnyel is rendelkezik (**lc()**, **rh()**, **apc()**, **cbd()**, **m6()**, **m7()**, **m8()**), megspórolva ezzel a szisztematikus komponens kézi bevitelét. Ezekhez a függvényekhez is hozzáadjuk a "logitP" és "cloglog" opciókat. A Plat modell definícióját a programkönyvtár dokumentációja [21, 13-14. oldal] tartalmazza. [lásd: standardModels.R, RHModel.R]

Modelldefinícióink "logitP" esetben a **StMoMo**-ban:

```

logitP_LC <- lc(link = "logitP")
logitP_RH <- rh(link = "logitP", cohortAgeFun = "1",
  approxConst = TRUE)
logitP_APC <- apc(link = "logitP")
f2 <- function(x, ages) x - mean(ages)
logitP_CBD <- StMoMo(link = "logitP", staticAgeFun = FALSE
  , periodAgeFun = c("1", f2))
logitP_M6 <- m6(link = "logitP")
logitP_M7 <- m7(link = "logitP")
f2 <- function(x, ages) mean(ages) - x
constPlat <- function(ax, bx, kt, b0x, gc, wxt, ages){...}
logit_PLAT <- StMoMo(link = "logit", staticAgeFun = TRUE,
  periodAgeFun = c("1", f2), cohortAgeFun = "1", constFun
  = constPlat)

```

A `constPlat` függvény a Plat modellre vonatkozó négy egyértelműsítő megkötelést írja le, amit a terjedelme miatt nem idézünk. A programkódból kiolvasható, hogy az RH modell esetében az egyszerűsített változatot definiáltuk ($\beta_x^{(0)} = 1$), és a Newton–Raphson módszerrel történő iteratív megoldást [10] választottuk.

Eddig csak definíciók szintjén bővítettük a programcsomagot; a modellek illesztése a `fit.StMoMo` függvénnyel történik. A "cloglog" esetben ez elég is, mivel az R beépített statisztikai csomagja tartalmazza a cloglog függvényt a *binomiális családban*, így ezt argumentumként megadva a kívánt modellt kapjuk. A logit függvény és a Poisson-eloszlás azonban olyan párosítás, amely nem képezi a beépített csomag részét, lehetőségünk van azonban bármilyen saját kapocsfüggvényt egyénileg definiálni (link-glm osztályú objektumként). Ez a definíció megtalálható Currie [5] cikkének függelékében. Hozzáadjuk a `StMoMo` programkönyvtár belső eszközkészletéhez. Ezután mint "logitP" függvény hívható meg, argumentumában a kitettséggel. [lásd: `fitStMoMo.R`, `internalUtils.R`]

A `fit.StMoMo` módosításainak lényegi részlete:

```

else if (object$link == "cloglog") {
  fittingModel <- gnm(formula = as.formula(object$
    gnmFormula),
    data = fitData, family= binomial(link = "cloglog"),
    weights = fitData$E * fitData$w, start = startCoef,
    verbose = verbose, ...)

```

```

} else if (object$link == "logitP") {
  fittingModel <- gnm(formula = as.formula(object$
    gnmFormula),
    data = fitData, family = poisson(link = logitP(fitData$E
      )),
    weights = fitData$w, start = startCoef,
    verbose = verbose, ...)
}

```

A két új esettel kibővítjük a log-likelihood, valamint a deviancia kiszámítását, és a hozzájuk tartozó figyelmeztetések listáját is.

Létrehozuk a vizsgált kor és időszak intervallumok vektorát, valamint az ezek méretének megfelelő `wxt` mátrixot, ami a három vagy annál kevesebb korévet tartalmazó kohorszokat eliminálja. Szükség esetén (binomiális modellek) a `central2initial()` függvénnyel áttérünk kezdeti kitettségre. (A HMD-ből letöltött adataink központi kitettséget tartalmaznak.) A `fit.StMoMo` függvénnyel végrehajtjuk az illesztést.

Példa illesztésre: logit-Poisson Lee-Carter és Renshaw-Haberman modellek. Az RH modellben az LC modell paramétereit adjuk meg kezdőértékként a gyorsabb konvergencia érdekében.

```

logitP_LCfit <- fit(logitP_LC, data = Data, ages.fit =
  ages.fit, years.fit = years.fit, wxt = wxt)
logitP_RHfit <- fit(logitP_RH, data = Data, ages.fit =
  ages.fit, years.fit = years.fit, wxt = wxt,
start.ax = logitP_LCfit$ax, start.bx = logitP_LCfit$bx,
  start.kt = logitP_LCfit$kt)

```

Ezzel `fitStMoMo` osztályú objektumokat kapunk. Az illesztett paramétereket a függelékben részletezett módon ábrázolhatjuk is.

5.3. Illeszkedésvizsgálat

Egy GAPC modell – és általában bármely halandósági modell – illeszkedését a *deviancia* fogalmának segítségével mérhetjük. Az egyes korcsoport-időszak kombinációkra érdemes kiszámolni az *egyedi skálázott devianciákat* (angolul *scaled deviance residuals*):

$$r_{x,t} = \text{sgn}(d_{x,t} - \mathbb{E}(D_{x,t})) \sqrt{\frac{\text{dev}_{x,t}}{\phi}}, \quad \text{ahol } \phi = \frac{\text{Dev}}{K - \nu},$$

valamint $K = \sum_{x=1}^{n_a} \sum_{t=1}^{n_y} \omega_{x,t}$ a megfigyelések, tehát a nem eliminált korcsoport-időszak kombinációk száma, ν pedig a modell effektív paramétereinek száma, lásd 4.2.7 táblázat. A $\omega_{x,t}$ mátrixot az előző szakaszban mint `wxt` definiáltuk a rövid kohorszok figyelmen kívül hagyására.

Továbbá az *egyedi devianciák* (ang. *deviance residuals*) képlete Poisson-eloszlás esetében

$$\text{dev}_{x,t} = 2 \left(d_{x,t} \log \frac{d_{x,t}}{E_{x,t}^c \hat{\mu}_{x,t}} + E_{x,t}^c \hat{\mu}_{x,t} - d_{x,t} \right),$$

míg binomiális eloszlás esetében

$$\text{dev}_{x,t} = 2 \left(d_{x,t} \log \frac{d_{x,t}}{E_{x,t}^0 \hat{q}_{x,t}} + (E_{x,t}^0 - d_{x,t}) \log \frac{E_{x,t}^0 - d_{x,t}}{E_{x,t}^0 (1 - \hat{q}_{x,t})} \right).$$

Ezek a képletek a 2. fejezetben bemutatott log-likelihood függvényekből és az egymásba ágyazott modellek² közötti választást segítő valószínűséghányados próba tesztstatistikájából származnak. [22]

A *teljes deviancia* (ang. *total deviance*):

$$\text{Dev} = \sum_{x=1}^{n_a} \sum_{t=1}^{n_y} \omega_{x,t} \text{dev}_{x,t}.$$

Ha az egyedi skálázott devianciákban fellelhető szabályos mintázat, az azt jelzi, hogy a modell nem képes az adatok összes jellemzőjét megfelelően leírni. [21] Ehhez a **StMoMo** csomagban található `plot` függvényekkel különböző típusú ábrákat készíthetünk. A devianciák kiszámításához a `residuals` függvényt leíró szkriptfájl is kibővítettük az új `cloglog` kapocsfüggvénnyel. [residualsfitStMoMo.R]

²Két modell egymásba ágyazott, ha egyik a másikból paraméter-megkötésekkel származtatható. Esetünkben a bővebb modell az ún. telített modell, amelyben a halálesetek számának elméleti várható értéke minden egyes korcsoport-időszak kombinációra megegyezik a tapasztalati értékkel, effektív paramétereinek száma tehát K . [22]

5.3.1. Futtatás a Currie által javasolt intervallumon

Ebben az esetben a teljes devianciákat vetjük össze különböző kapocsfüggvények alkalmazása mellett. Példaként bemutatjuk az Egyesült Államokra vonatkozó táblázatot.

Modell	$\log \mu, \mathcal{P}$	$\text{logit} \mu, \mathcal{P}$	$\text{logit} q, \mathcal{B}$	$\text{cloglog} q, \mathcal{B}$
LC	41557	403036	41331	106585
RH	11247	11121	11197	11259
APC	26216	393853	25023	90906
CBD	75860	371986	91210	149830
M6	30249	375026	34253	97913
M7	15671	380035	15015	80598
Plat	13004	376392	13008	78387

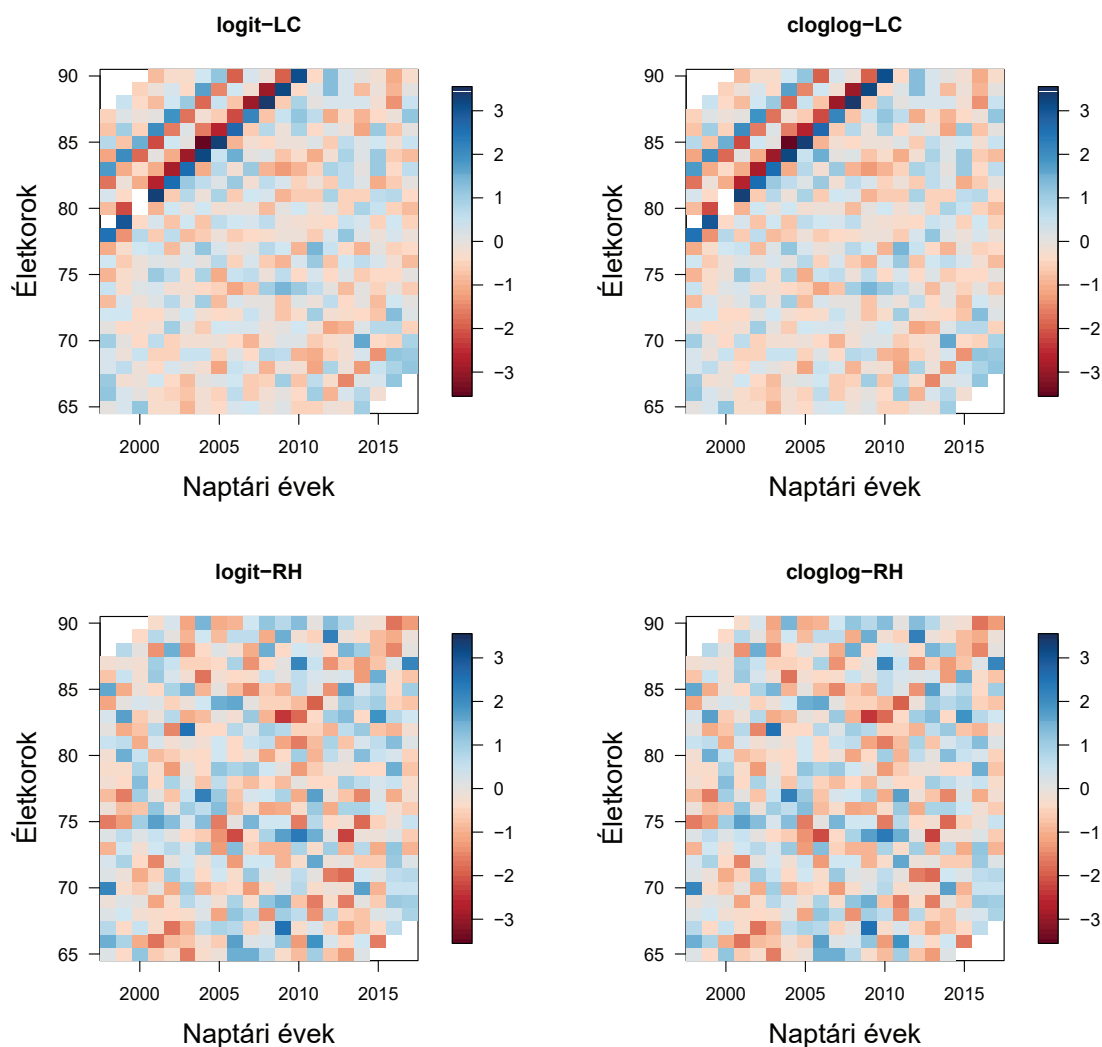
Egyesült Államok, 1960–2009., 50–90 éves férfi népesség.

A táblázatból azt látjuk, hogy a **StMoMo** az új kapocsfüggvényekkel futtatva minden **gnm** illesztéssel operáló modellre rosszabb eredményt ad, mint a kanonikus függvénnyel. Ez ellentmond várakozásainknak. Fontos megjegyezni, hogy bár a modellek egymás közötti rangsorrendje a legtöbb esetben megegyezik a Currie cikében [5] látottakkal, a **StMoMo** ugyanazon bemeneti paraméterekre eltérő teljes devianciákat mutat. Az eltérés minden ország esetében jelen van, ezért lehetséges, de nem valószínű az adatok esetleges frissítéséből adódó különbség. Ezek alapján a GAPC modellesaládot implementáló **StMoMo** programcsomag működése és bővítési lehetőségei további, dolgozatunkon túlmutató célirányos elemzést igényelnek.

5.3.2. Futtatás saját intervallumon

Saját, 5.1. pontban leírt intervallumunk esetében az egyedi skálázott devianciák ábrázolására is mutatunk példát. A következő ábrákon a binomiális eloszlás esetét láthatjuk logit és cloglog kapocsfüggvényekkel, a Lee–Carter és a Renshaw–Haberman modellekre. Szembetűnő, hogy a kohorszhatás hozzáadása hogyan tünteti el az 1915–1920-as kohorszok kiugró értékeit mindkét kapocsfüggvény esetében.

Az általunk választott intervallumról ugyanaz állapítható meg, amit a Currie által javasolt esetén leírtunk. Az intervallum változtatása nem okozott lényeges különbséget, ezért ez is kizárható, mint az eltérések oka.



5.1. ábra. Egyedi skálázott devianciák Lee–Carte és Renshaw–Haberman modellekre. Binomiális eloszlás mellett a bal oldalon logit, a jobb oldalon cloglog kapocsfüggvényt feltételezve. Uniszex magyar adatok a saját intervallumon.

5.4. Előrejelzések

A modellek illesztése természetesen csak az első lépés; az aktuáriusi szempontból érdekes kérdések megválaszolásához a cél ezekkel a modellekkel előrejelzéseket készíteni az élettartam-kockázat alakulására. Az eltérő eredmények okának feltárása után a következő feladat a bővítés kiterjesztése a programcsomag vonatkozó részeire, és az ezzel kapcsolatos kérdések elemzése. Jobb illesztést kínáló kapocsfüggvény vajon jobb előrejelzést is kínál? A válasz nem egyértelmű, mert pontosabban illeszkedő modellből még nem következik, hogy pontosabb előrejelzésre is alkalmas. Ha mégis így van, érdemes megvizsgálni mely kapocsfüggvényekkel készíthetők pontosabb előrejelzések.

6. fejezet

Összefoglalás

Dolgozatunkban bemutattunk halandóságot leíró mérőszámokat, és ezek előrejelzésére irányuló sztochasztikus modelleket. Bemutattunk több ismert modellt és ezeket az általánosított korcspot–időszak–kohorsz (GAPC) model családj keretébe foglalva rávilágítottunk azok hasonlóságaira és különbségeire. Currie 2016-ban megjelent cikke [5] alapján felvetettük a kérdést, hogy különböző kapcsolfüggvények alkalmazásával javítható-e bármelyik vizsgált modell illeszkedése a magyar vagy lengyel halálozási adatokra.

Hogy az összehasonlítás egyszerűbb, könnyen paraméterezhető és reprodukálható legyen, az R programcsomag a GAPC elméletén alapuló **StMoMo** programkönyvtárát [21] szeretnénk volna használni. A csomagban a model családba tartozó modellek rendelkezésünkre álltak, azonban két javasolt, a megszokottól eltérő kapcsolfüggvény hiányzott.

Az említett célok érdekében a nyílt forráskódú programkönyvtárat kibővítettük a két hiányzó elemmel. A kapcsolfüggvények definiálásán kívül elvégeztük minden, a modellek illesztéséhez és az illeszkedésvizsgálathoz szükséges függvény frissítését. Az illeszkedés vizsgálata során a Currie által publikáltaktól eltérő eredményt kaptunk. A kibővített **StMoMo**val végzett számítások szerint az általánostól eltérő kapcsolfüggvények egy esetben sem adtak jobb eredményt, mint a kanonikus változat. Az eltérések oka további vizsgálatot igényel.

A fő kérdésen túlmutatóan folyamatban van az előrejelzésekhez tartozó függvények kiterjesztése is. Felvettük a kapcsolatot a programkönyvtár szerzőjével, és reméljük, hogy ha elkészültünk, munkánk része lehet a nyilvános verzióknak.

7. fejezet

Függelék

A StMoMo módosított szkriptfájljainak listája

- fitStMoMo.R
- residualsfitStMoMo.R
- StMoMo.R
- fittedfitStMoMo.R
- RHModel.R
- internalUtils.R
- standardModels.R

Az egyedi skálázott devianciák ábrázolása

A maradékok ábrázolására a **StMoMo** háromféle lehetőséget kínál.

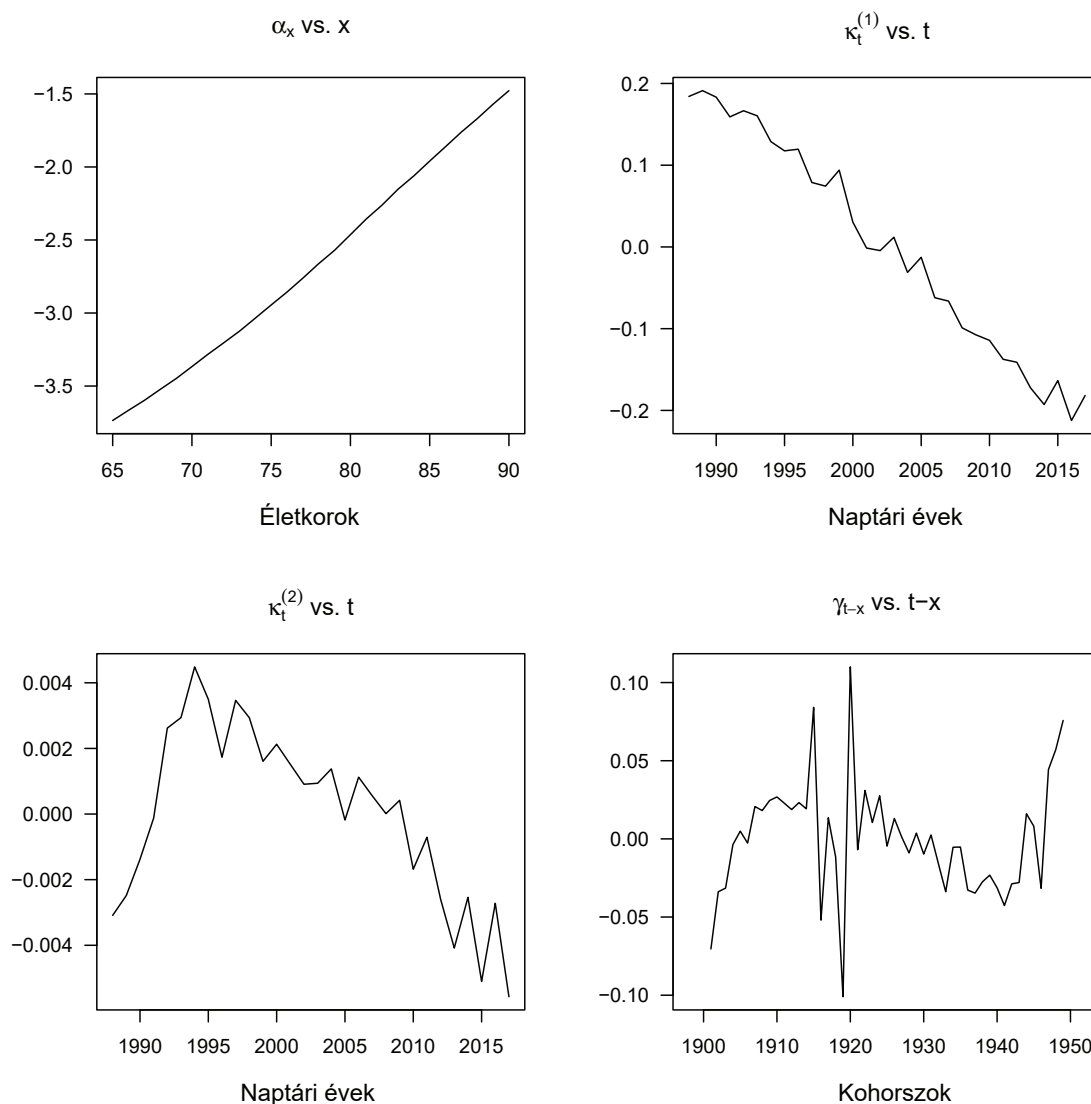
```
plot(cloglog_LCres, type = "colourmap", reslim = c(-3.5,
  3.5), main = "cloglog_LC")
plot(cloglog_LCres, type = "scatter", reslim = c(-3.5,
  3.5), main = "cloglog_LC")
plot(cloglog_LCres, type = "signplot", reslim = c(-3.5,
  3.5), main = "cloglog_LC")
```

A `scatter` ábrával a modellek egy dimenziós szisztematikus hibáját ellenőrizhetjük, de ez bizonyos kereszthatásokat eltakarhat. Kereszthatások kimutatására a másik két típus alkalmas. A `colourmap`-re példa az 5.1. ábra. A `signplot` a `colourmap` egy durvább változata.

A becsült paraméterek ábrázolása

Egyszerűen alkalmazhatjuk a `plot` függvényt `fitStMoMo` osztályú objektumokra, több becsült paraméter esetén több oszlopban (jelen esetben kettő) megjelenítve, illetve szükség esetén jelezve, hogy a modellben szereplő β_x nem paraméteres.

```
plot(log_PLATfit, parametricbx = FALSE, nCol = 2, las = 1,
      cex.lab = 1.2)
```



7.1. ábra. Példaként bemutatjuk a log-Poisson Plat modellel illesztett paramétereket. A γ_{t-x} kohorsz-paraméter 1915–1920-as kiugró értékei az első világháborús generációk eltérő mortalitási jellemzőit kompenzálja.

Uniszex magyar adatok, 1988–2017., 65–90 életkorú népesség.

Irodalomjegyzék

- [1] Natacha Brouhns, Michel Denuit, and Jeroen K Vermunt. A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373–393, 2002.
- [2] Andrew JG Cairns, David Blake, and Kevin Dowd. A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, 73(4):687–718, 2006.
- [3] Andrew JG Cairns, David Blake, Kevin Dowd, Guy D Coughlan, David Epstein, Alen Ong, and Igor Balevich. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35, 2009.
- [4] Iain D Currie. Smoothing and forecasting mortality rates with P-splines. *Talk given at the Institute of Actuaries*, 2006.
- [5] Iain D Currie. On fitting generalized linear and non-linear models of mortality. *Scandinavian Actuarial Journal*, 2016(4):356–383, 2016.
- [6] Agnieszka Fihel and Domantas Jasilionis. About Mortality Data for Poland. <https://mortality.org/hmd/POL/InputDB/POLcom.pdf>. Human Mortality Database, 2021., letöltés dátuma: 2021. 05. 17.
- [7] Benjamin Gompertz. XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. FRS &c. *Philosophical Transactions of the Royal Society of London*, (115):513–583, 1825.
- [8] Steven Haberman and Arthur Renshaw. A comparative study of parametric mortality projection models. *Insurance: Mathematics and Economics*, 48(1):35–55, 2011.

- [9] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). <https://mortality.org>. Az adatok letöltésének dátuma: 2020. 12. 14. - 2021. 05. 23.
- [10] Andrew Hunt and David P Blake. On the Structure and Classification of Mortality Models. Technical report, PI-1506, Pensions Institute, 2015.
- [11] D. Jdanov D.A. Gleijer J.R. Wilmoth, K. Andreev and D. Philipov V. Shkolnikov P. Vachon C. Winant M. Barbieri T. Riffe with the assistance of C. Boe, M. Bubenheim. Methods Protocol for the Human Mortality Database. <https://mortality.org/Public/Docs/MethodsProtocol.pdf>, 2021. (Rövid összefoglaló elérhető: <https://mortality.org/Public/Docs/MP-Summary.pdf>) Letöltés dátuma: 2021. 05. 23.
- [12] Kolos Cs. Ágoston and Erzsébet Kovács. Halandósági modellek. <https://www.uni-corvinus.hu/alfresco/dokumentumtar/download/?id=f097c0d3-792b-4507-a9c0-426f8aa6826f;1.1>. Aktuárius jegyzetek, 3. kötet, 2000., letöltés dátuma: 2021. 05. 13.
- [13] Ronald D Lee and Lawrence R Carter. Modeling and forecasting US mortality. *Journal of the American Statistical Association*, 87(419):659–671, 1992.
- [14] László Németh, Domantas Jasilionis, László Radnóti. About Mortality Data for Hungary. <https://mortality.org/hmd/HUN/InputDB/HUNcom.pdf>. Human Mortality Database, 2018., letöltés dátuma: 2021. 04. 23.
- [15] Richard Plat. On stochastic mortality modeling. *Insurance: Mathematics and Economics*, 45(3):393–404, 2009.
- [16] R Core Team. R: A language and environment for statistical computing. <https://www.R-project.org/>, 2020. R Foundation for Statistical Computing, Vienna, Austria.
- [17] László Radnóti. Az élettartamok statisztikája. *Statisztikai Szemle*, 7:559–570, 2003.
- [18] Arthur E Renshaw and Steven Haberman. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, 38(3):556–570, 2006.
- [19] AJ Russell. Life Contingencies. By Alistair Neill [Hutchinson, 1977]. *Journal of the Institute of Actuaries*, 105(2):209–210, 1978.

- [20] Heather Turner and David Firth. *Generalized nonlinear models in R: An overview of the gnm package*, 2020. R package version 1.1-1.
- [21] Andrés M. Villegas, Vladimir K. Kaishev, and Pietro Millossovich. StMoMo: An R Package for Stochastic Mortality Modeling. *Journal of Statistical Software*, 84(3):1–38, 2018.
- [22] Péter Vékás. *Az élettartam-kockázat modellezése*. PhD thesis, Corvinus University of Budapest, 2016.
- [23] Rob J Hyndman with contributions from Heather Booth, Leonie Tickle, and John Maindonald. *demography: Forecasting Mortality, Fertility, Migration and Population Data*, 2019. R package version 1.22.