Eötvös Loránd Tudományegyetem
Természettudományi kar

Budapesti Corvinus Egyetem
Közgazdaságtudományi Kar

Készítette: Török Anikó

Túlélési modellek a biztosításban

Biztosítási és Pénzügyi Matematika MSc

Aktuárius szakirány

Szakszemináriumvezető: Vékás Péter

Operációkutatás és Aktuáriustudományok Tanszék

Budapest, 2018

Köszönetnyilvánítás

Szeretnék köszönetet mondani témavezetőmnek, Vékás Péternek, hogy mindig a rendelkezésemre állt, szakmai tanácsaival és segítőkészségével nagyban segítette ezen dolgozat elkészültét.

Köszönettel tartozom továbbá családomnak és barátaimnak támogatásukért és belém vetett bizalmukért, nélkülük nem jutottam volna el idáig.

Végül, de nem utolsó sorban szeretném megköszönni barátomnak, Kiss Kolosnak az ebben az időszakban tanúsított végtelen türelmét, és elgondolkoztató hozzászólásait a dolgozatom témájához.

Tartalomjegyzék

1. fejezet

Bevezetés

Szakdolgozatom célja elsősorban a túlélési modellek biztosításban való alkalmazhatóságának vizsgálata. Biztosítási kalkulációkban számos helyen alkalmazhatóak a túlélési modellek, ebben a dolgozatban egy speciális területre koncentráltam. Segítségükkel elsősorban házaspárok élettartamának egymással való összefüggését vizsgáltam. Az életbiztosítási számításokban legtöbbször egymástól függetlennek tekintik egy házaspár két tagjának várható élettartamát, ám ez nem feltétlenül állja meg a helyét a valóságban.

Tekintettel arra, hogy a férj és a feleség azonos életkörülmények között élnek, hasonló szokásokkal rendelkeznek és a pénzügyi hátterük is megegyezik, jogosan feltételezhetjük, hogy élettartamuk nem független egymástól. Ám a közös életvitelen felül a partner halála is jelentős befolyással lehet a halandósági valószínűségre. Ennek az eseménynek a rövid-, illetve hosszútávú hatásait fogom vizsgálni különböző módszerekkel.

A következő fejezet egy túlélési analízisről szóló rövid matematikai bevezetés. Az itt végigvett fogalmak kerülnek felhasználásra a 3. fejezetben, ahol is részletesebben bemutatom a két, ebben a témakörben legelterjedtebb eljárást, a Kaplan-Meier modellt és a Cox regressziót.

A 4. fejezetben az egymással összefüggő élettartamok vizsgálatának egy fontos részletét mutatom be, név szerint a frailty modelleket. Ha bizonyos valószínűségi változók - például élettartamok - közötti kapcsolatot szeretnénk modellezni, mindenképp szót kell

ejtenünk az egyik legelterjedtebb módszerről, a kopulákról. Az 5. fejezetben kerülnek bemutatásra ennek a módszernek az alapfogalmai, illetve egy-két szélesebb körben használt nevezetes kopula. A 6. fejezetben a két modellezési alapötlet közötti hasonlóságokat és különbségeket tárom fel.

A 7. fejezetben egy konkrét adatbázison modellezem az eddig tárgyalt összefüggést az élettartamok között, azon belül is a házastárs halálának hatását az életben maradt tag túlélési valószínűségére. Az elemzéshez három részre bontottam a házaspárok élettartamait. Először azt a fázist vizsgáltam, amikor mindketten életben vannak, majd az első halál időpontjától vett rövid intervallumot, végül pedig a fennmaradó intervallumot (ha volt ilyen).

Így lehetőségem nyílt megvizsgálni a túlélési valószínűségek alakulását abban az időszakban, mikor még egyik félnek sem kellett elviselnie párja halálát, röviddel a haláleset után, illetve hosszútávon. Vizsgálódtam a nemek tekintetében is: máshogy hat-e a férfiakra és a nőkre társuk elvesztése? Különbözik-e a társ halálának hatása rövid-, és hosszútávon? A kapott ereményeim a 8. fejezetben foglalom össze, és további elemzési és javítási lehetőségeket vetek fel.

2. fejezet

Alapfogalmak

A túlélési modellek alapvető célja egy adott eseményig eltelt idő eloszlásának modellezése. Jellemző ilyen időpont például a halálozásig eltelt idő, innen ered a modell elnevezése is. De természetesen alkalmazható más időtartamok modellezésére is, például gyakran használják törléselemzésre is, azaz biztosítási szerződések megkötésétől a törlésig eltelt idő modellezésére.

Természetesen ez az adott esemény lehet pozitív dolog is, mint például egy betegség diagnosztizálásától a gyógyulásig eltelt idő. Hasonlóképpen használták már az orvostudmányban egy adott betegség visszatérési idejének modelllezésére is.

Most nézzünk néhány matematikai fogalmat a megértéshez (lásd [7] és [3]).

2.1. Egyváltozós eset

2.1. Definíció. *Túlélésfüggvény alatt értjük az*

$$S(t) = P(T \geq t)$$

valószínűséget, ahol $t > 0$, $T > 0$ *pedig egy valószínűségi változó, úgynevezett túlélési idő.*

Látható, hogy a túlélésfüggvény előállítható az eloszlásfüggvényből, hiszen

$$S(t) = 1 - F(t).$$

Tehát annak a valószínűségét jelöli, hogy az adott esemény egy bizonyos idő előtt nem következett be. Azaz, ha T élettartamot jelöl, akkor $S(t)$ annak a valószínűsége, hogy az alany megélte a t időpontot. Használatos még a túlélésfüggvény egyenlőséget nem megengedő változata is, azaz amikor az alany túlélte a t időpontot. A később ismertetett túlélési modellek egyik célja bizonyos adatokból ennek a túlélési görbe alakjának megbecslése.

A túlélési modellek elméletében a következő legfontosabb fogalom a hazárdráta, másnéven kockázati ráta, amely például a később bemutatott Cox-regresszióban játszik fontos szerepet, ugyanis ott a túlélésfüggvény helyett a hazárdráta logaritmusát becsüljük a magyarázóváltozók segítségével.

2.2. Definíció. *A hazárdrátát a következő határérték adja meg:*

$$\lambda(t) = \lim_{dt \to 0+} \frac{P(T < t + dt | T \geq t)}{dt}$$

A hazárdráta jelöli annak a valószínűségét, hogy az egyed egy adott időpillanatban rövid időn belül (dt) meghal, úgy hogy az adott időpillanatban még életben volt.

A hazárdráta felírható a sűrűség-, illetve a túlélésfüggvény hányadosaként is (ha a túlélésfüggvény nem 0), ugyanis a feltételes valószínűség tulajdonságát felhasználva

$$\begin{aligned} \lambda(t) &= \lim_{dt \to 0+} \frac{P(T < t + dt | T \geq t)}{dt} = \lim_{dt \to 0+} \frac{P(t \leq T < t + dt)}{P(T \geq t) dt} \\ &= \lim_{dt \to 0+} \frac{F(t + dt) - F(t)}{S(t) dt} = \frac{1}{S(t)} \lim_{dt \to 0+} \frac{F(t + dt) - F(t)}{dt} \\ &= \frac{F'(t)}{S(t)} = \frac{f(t)}{S(t)}, \end{aligned} \tag{2.1.1}$$

ahol f sűrűség-, F az eloszlás-, $S \neq 0$ pedig a túlélésfüggvény.

A későbbi alkalmazásokhoz vezessük be a kumulált kockázati ráta definícióját is.

2.3. Definíció. *A kumulált kockázati rátát az alábbi integrál adja meg:*

$$\Lambda(t) = \int_0^t \lambda(s) ds.$$

Ennek a definíciónak a későbbi fejezetben tárgyalt Cox-regressziónál lesz fontos szerepe.

2.2. Kétváltozós eset

Az alábbi fogalmak kiterjeszthetők két, illetve több változóra is, amelyek használatosak lehetnek a több változó (például élettartamok) közötti összefüggés modellezésére.

2.4. Definíció. *Együttes túlélésfüggvény alatt az*

$$S_{12}(t_1, t_2) = P(T_1 \geq t_1, T_2 \geq t_2)$$

valószínűséget értjük, ahol $t_1, t_2 > 0$*,* T_1, T_2 *pedig túlélési idők.*

A korábbiakhoz hasonlóan $S_{12}(t, t)$ annak a valószínűségét jelenti, hogy mindkét egyed életben van t időpontban.

Felírhatóak természetesen a marginális túlélésfüggvények, például

$$S_1(t_1) = P(T_1 \geq t_1) = S_{12}(t_1, 0),$$

amelyek az előbb bemutatott egyváltozós túlélésfüggvények tulajdonságaival rendelkeznek. Ha a két túlélési idő független lenne egymástól, akkor az együttes túlélésfüggvény a marginálisok szorzata lenne.

2.5. Definíció. *Az alábbi módon definiálható a feltételes túlélésfüggvény:*

$$S_{1|2}(t_1|T_2 = t_2) = P(T_1 \geq t_1|T_2 = t_2),$$

ahol $t_1, t_2 > 0$*,* T_1, T_2 *túlélési idők.*

Ez a fajta feltételes túlélési függvény annak a valószínűségét adja meg, hogy az első egyed életben van t_1-ben, feltéve, hogy a másik egyed t_2-ben halt meg. Használatos még az egyenlőség helyett $\geq$-t írni, ekkor a feltétel értelemszerűen arra módosul, hogy a másik egyed még élt t_2-ben.

Nézzük most a kétváltozós hazárdrátát, amely logikus kiterjesztése az egyváltozós esetnek.

2.6. Definíció. *Az együttes hazárdráta az z alábbi módon definiálható:*

$$\lambda_{12}(t_1, t_2) = \lim_{dt \to 0} \frac{P(T_1 < t_1 + dt, T_2 < t_2 + dt | T_1 \geq t_1, T_2 \geq t_2)}{dt^2}$$

A marginális hazárdráta pedig az alábbi alakot veszi fel:

$$\lambda_1(t_1) = \lim_{dt \to 0} \frac{P(T_1 < t_1 + dt | T_1 \geq t_1)}{dt}$$

Látható, hogy a túlélési idők függetlensége esetén az együttes hazárdráta a marginálisok szorzata, ugyanis ebben az esetben

$$\lambda_{12}(t_1, t_2) = \lim_{dt \to 0} \frac{P(T_1 \in [t_1, t_1 + dt), T_2 \in [t_2, t_2 + dt))}{P(T_1 \geq t_1, T_2 \geq t_2) dt^2}.$$

Tehát a függetlenség miatt

$$\begin{aligned} \lambda_{12}(t_1, t_2) &= \lim_{dt \to 0} \frac{P(T_1 \in [t_1, t_1 + dt)) P(T_2 \in [t_2, t_2 + dt))}{P(T_1 \geq t_1) P(T_2 \geq t_2) dt^2} \\ &= \lim_{dt \to 0} \frac{P(T_1 < t_1 + dt | T_1 \geq t_1)}{dt} \frac{P(T_2 < t_2 + dt | T_2 \geq t_2)}{dt} \\ &= \lambda_1(t_1) \lambda_2(t_2) \end{aligned} \tag{2.2.2}$$

2.7. Definíció. *Feltételes hazárdrátának nevezzük az alábbi kifejezést:*

$$\lambda_{1|2}(t_1 | T_2 = t_2) = \lim_{dt \to 0} \frac{P(T_1 < t_1 + dt | T_1 \geq t_1, T_2 = t_2)}{dt}.$$

Ennél a kifejezésnél is használatos egyenlőség helyett $\geq$-t írni, azaz:

$$\lambda_{1|2}(t_1 | T_2 \geq t_2) = \lim_{dt \to 0} \frac{P(T_1 < t_1 + dt | T_1 \geq t_1, T_2 \geq t_2)}{dt}.$$

A túlélési modellek céljai (lásd [3]) elsősorban az eddig bemutatott fogalmak becslése és értelmezése, azaz

- A túlélésfüggvények és/vagy hazárdráták becslése a rendelkezésre álló adatokból
- Ezen függvények értelmezése, összehasonlítása
- A magyarázó változók és a túlélési idők közti kapcsolat feltárása

3. fejezet

Modellek

Ebben a fejezetben a túlélés analízis két leggyakrabban használt modelljét fogom bemutatni, és összehasonlítani. A Kaplan-Meier modellt egyszerűsége, a Cox regressziót pedig összetettebb értelmezhetősége miatt alkalmazzák széles körökben (lásd [3]).

3.1. Kaplan-Meier modell

A Kaplan-Meier modell a legelterjedtebb nemparaméteres túlélési modell. Ilyenkor nem alkalmazunk magyarázóváltozókat a modellben, mindössze a túlélésfüggvényt próbáljuk megbecsülni.. A modellhez tekintsük a vizsgált eseményekig (például halál) eltelt időket. Jelöljük t_f-fel egy adott esemény bekövetkezéséig eltelt időt (failure time), amit néhol kilépési időnek neveznek, mivel ekkor lép ki a megfigyelés a modellünkből. Rendezzük ezeket a kilépési időket növekvő sorrendbe.

Először is tekintsük a második fejezetben tárgyalt túlélésfüggvénynek egy másik verzióját, amelyben nem megengedett az egyenlőség, tehát

$$S_+(t) = P(T > t),$$

ahol $t > 0$ pozitív szám, $T > 0$ valószínűségi változó.

A Kaplan-Meier modell az alábbi formulával becsli a módosított túlélési függvényt

ebben a pontban:

$$\hat{S_+}(t_f) = \hat{S_+}(t_{f-1})P(T > t_f | T \geq t_f), \tag{3.1.1}$$

ahol $\hat{S_+}$ jelöli a modell által megbecsült túlélési függvényt.

Azaz a becslés egy rekurzív szorzat formájában áll elő, ahol az egyik tag az előző eseményig eltelt kilépési idő (t_{f-1}) túlélésének a valószínűsége, a másik tag pedig t_f túlélésének a valószínűsége, feltéve hogy legalább t_f-ig életben volt az egyed.

A fenti képletbe $\hat{S_+}(t_{f-1})$ helyébe behelyettesítve a

$$\hat{S_+}(t_{f-1}) = \prod_{i=1}^{f-1} P(T > t_i | T \geq t_i)$$

képletet, megkapjuk a

$$\hat{S_+}(t_f) = \prod_{i=1}^{f} P(T > t_i | T \geq t_i)$$

formulát.

De miért épp (3.1.1) egyenletet használjuk a túlélésfüggvény becslésére? Miért igaz ez az eredeti túlélésfüggvényre? A megértéshez tekintsük az alábbi két eseményt:

$$A := \{T \geq t_f\},$$

$$B := \{T > t_f\}.$$

Felhasználva a

$$P(A \cap B) = P(A)P(B|A) \tag{3.1.2}$$

tulajdonságát a metszet valószínűségének, azonnal adódik az egyenlet, ugyanis esetünkben

– $P(A \cap B) = P(T \geq t_f \cap T > t_f) = P(T > t_f) = S_+(t_f)$,

– $P(A) = P(T \geq t_f) = P(T > t_{f-1})$,

– $P(B|A) = P(T > t_f | T \geq t_f)$,

tehát (3.1.2) miatt

$$S_+(t_f) = S_+(t_{f-1})P(T > t_f | T \geq t_{f-1}),$$

ahol t_f kilépési idő.

3.2. Cox regresszió

A Kaplan-Meier modell mellett a leggyakrabban alkalmazott eszköz a túlélés analízis során a Cox regresszió. Előnye az előbbivel szemben, hogy paraméteres modell, azaz bizonyos magyarázó változók hatásai is vizsgálhatóak a modellben.

A túlélésfüggvény helyett ez az eljárás a 2. fejezetben leírt hazárdrátát - illetve annak logaritmusát - becsli először, az alábbi egyenlettel:

$$\lambda(t) = \lambda_0(t)e^{\sum_{i=1}^{p} \beta_i x_i}, \tag{3.2.3}$$

ahol $t > 0$, és az x_i-k $(i = 1 \dots p)$ a magyarázóváltozók. Valamint λ_0-lal a baseline hazárdrátát jelöljük, azaz azt az esetet amikor egy megfigyelésre minden magyarázóváltozó 0 értéket vesz fel.

Látható ebből, hogy

$$\ln \lambda(t) = \ln \lambda_0(t) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

De hogy kaphatjuk meg ebből a túlélésfüggvényt? A korábban tárgyalt kumulált kockázati ráta segítségével, ugyanis (2.1.1) miatt

$$\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^t \frac{f(s)}{S(s)}ds = [-\ln S(s)]_0^t = -\ln S(t) + \ln S(0) = -\ln S(t),$$

emiatt

$$S(t) = e^{-\Lambda(t)}. \tag{3.2.4}$$

Továbbá visszahelyettesítve az előbb megkapott hazárdráta becslést a kumulált hazárdráta képletébe:

$$\Lambda(t) = \int_0^t \lambda(s)ds = \int_0^t \lambda_0(t)e^{\sum_{i=1}^{p} \beta_i x_i} = \Lambda_0(t)e^{\sum_{i=1}^{p} \beta_i x_i},$$

ahol $t > 0$, és $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$

Tehát (3.2.4) alapján

$$S(t) = \exp(-\Lambda_0(t)e^{\sum_{i=1}^{p} \beta_i x_i}).$$

4. fejezet

Frailty modellek

A frailty modellek a túlélésfüggvények elméletének egy fontos részét alkotják (lásd [7], [2] és [8]). A frailty szó magyarul törékenységet, gyengeséget jelent. Az általános túlélési modellek homogenitást feltételeznek, azaz mindenki ugyanazzal a hazárd-, illetve túlélési függvénnyel rendelkezik, ám ez a valóságban nincs mindig így. A frailty modell mögötti elgondolás az, hogy létezik egy véletlenszerű hatás, ami felelős a heterogenitásért, ezt nevezzük frailty-nek.

A modell alapötlete az, hogy a túlélési idők hátterében két ok áll, először is a véletlen, amit a hazárdráta ír le. Másodszor pedig a frailty, ami lehet közös több egyednél is, ez írja le a köztük lévő összefüggést. Például testvérek, vagy házaspárok túlélési ideje (élethossza) nem független egymástól. A fent említett frailty-t, azaz a véletlenszerű hatást nevezzük jelöljük ezentúl Θ-val.

4.1. Egyváltozós frailty modell

A modellt leíró feltételes hazárdráta $\Theta = \theta$ és $t > 0$ esetén az alábbi:

$$\lambda(t|\theta) = \theta\lambda_0(t),$$

ahol λ_0 a baseline hazárdráta.

A frailty egy nem megfigyelt véletlen változó, amely növeli az egyed kockázatát, ha nagyobb, illetve csökkenti, ha kisebb. Eloszlásáról, ezáltal a sűrűséggfüggvényéről is rendelkezünk feltevéssel.

A modell feltételes túlélésfüggvénye természetesen felírható a kumulált hazárdrátával

$$S(t|\theta) = e^{-\theta\Lambda_0(t)}$$

alakban. Ebből a feltétel nélküli túlélés függvényt az alábbi egyenlet adja meg:

$$S(t) = \int S(t|\theta)g(\theta)d\theta = \int e^{-\theta\Lambda_0(t)}g(\theta)d\theta,$$

ahol g a Θ valószínűségi változó sűrűségfüggvénye.

Látható, hogy ez Θ Laplace transzformáltja $s = \Lambda_0(t)$ helyen, azaz

$$S(t) = \mathcal{L}_\Theta(\Lambda_0(t))$$

A sűrűségfüggvény a túlélők között, azaz azon egyedek között, akiknek a túlélési ideje nagyobb, mint a jelen, így alakul:

$$g(\theta|T \geq t) = \frac{P(T \geq t|\theta)g(\theta)}{P(T \geq t)} = \frac{S(t|\theta)g(\theta)}{\int_0^\infty S(t|\theta)g(\theta)d\theta} = \frac{S_0(t)^\theta g(\theta)}{\int_0^\infty S_0(t)^\theta g(\theta)d\theta},$$

ahol S_0 a baseline túlélésfüggvény, azaz $S_0(t) = e^{-\Lambda_0(t)}$.

Ezt, illetve a várható érték definícióját felhasználva a frailty várható értéke a túlélő egyedekre:

$$E(\theta|T \geq t) = \int_0^\infty \theta g(\theta|T \geq t)d\theta = \frac{\int_0^\infty \theta S_0(t)^\theta g(\theta)d\theta}{\int_0^\infty S_0(t)^\theta g(\theta)d\theta}.$$

4.2. Kétváltozós frailty modell

Tegyük fel, hogy létezik egy olyan θ, hogy T_1 és T_2 feltételesen függetlenek θ feltétellel.

T_1 és T_2 feltételes függetlensége miatt az együttes túlélésfüggvény a marginális túlélésfüggvények szorzata a feltételes esetben, azaz

$$S_{12}(t_1, t_2|\theta) = S_1(t_1|\theta)S_2(t_2|\theta).$$

Továbbá feltesszük, hogy

$$S_i(t_i|\theta) = S_{0i}(t_i)^\theta,$$

ahol S_{0i} baseline túlélési függvény (ami akkor áll elő, ha minden magyarázó változó nulla).

Így a feltételes kumulált hazárdráták felírhatóak

$$\begin{aligned} \Lambda_i(t_i) := \Lambda_i(t_i|\theta) &= -\ln S_i(t_i|\theta) = -\ln S_{0i}(t_i)^\theta \\ &= \theta(-\ln S_{0i}(t_i)) = \theta\Lambda_{0i}(t_i) \end{aligned} \tag{4.2.1}$$

alakban, felhasználva a túlélésfüggvény és a kumulált hazárdráta közötti kapcsolatot. Tehát a véletlen hatás multiplikatívan hat a hazárdrátára.

Emiatt a feltételes együttes túlélésfüggvény alakja a következő:

$$S_{12}(t_1, t_2|\theta) = S_{01}(t_1)^\theta S_{02}(t_2)^\theta = e^{-\theta\Lambda_{01}(t_1)}e^{-\theta\Lambda_{02}(t_2)},$$

felhasználva, hogy $S_{0i}(t_i) = e^{-\Lambda_{0i}(t_i)}$.

Tehát

$$S_{12}(t_1, t_2|\theta) = e^{-\theta(\Lambda_{01}(t_1)\Lambda_{02}(t_2))}. \tag{4.2.2}$$

Ahhoz, hogy a feltétel nélküli együttes túlélési függvényt megkapjuk, ki kell integrálnunk a fenti egyenletből θ-t.

A feltételes eloszlás, ezáltal a feltételes túlélési függvény tulajdonságait ismerve látható, hogy

$$S_{12}(t_1, t_2) = \int_0^\infty S_{12}(t_1, t_2|\theta)g(\theta)d\theta.$$

Ide visszahelyettesítve (4.2.2)-t, megkapjuk, hogy

$$S_{12}(t_1, t_2) = \int_0^\infty e^{-\theta(\Lambda_{01}(t_1)\Lambda_{02}(t_2))} g(\theta) d\theta.$$

Látható, hogy ez g Laplace transzformáltja $s = \Lambda_{01}(t_1)\Lambda_{02}(t_2)$ helyen, vagyis

$$S_{12}(t_1, t_2) = \mathcal{L}_g(\Lambda_{01}(t_1)\Lambda_{02}(t_2)). \tag{4.2.3}$$

Mivel használtuk Θ sűrűségfüggvényét, nyilván volt egy feltevésünk az eloszlására. A heterogenitást gyakran modellezik exponenciálisak összegével, ezért a Θ frailty sokszor gamma eloszlású. Felhasználva a gamma eloszlás Laplace transzformáltját, a kétváltozós túlélésfüggvény az alábbi fogja alakot ölteni:

$$S_{12}(t_1, t_2) = (S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{-\frac{1}{\sigma^2}}. \tag{4.2.4}$$

5. fejezet

A kopula függvények

5.1. Korrelációs együtthatók

Ha két valószínűségi változó közötti kapcsolatot akarjuk vizsgálni, mindenképp szót érdemelnek a korrelációs együtthatók, illetve a kopula függvények (lásd [11] és [4]). Az összefüggés leírásának legegyszerűbb módja a korrelációs együtthatók használata. Az első, és egyben legelterjedtebb a lineáris korreláció, másnéven Pearson korreláció, amely ismerve a két változó eloszlását az alábbi képlet alapján adódik:

$$R(X,Y) = \frac{cov(X,Y)}{D(X)D(Y)}.$$

A lineáris korreláció értéke -1 és 1 között van, és akkor 1, ha a változók között tökéletes pozitív, akkor -1, ha tökéletes negatív lineáris kapcsolat van. Ezt a feltételt gyengíti Spearman mutatója, a rangkorreláció, amely szintén -1 és 1 között veszi fel az értékeit:

$$\rho(X,Y) = R(F_X(X), F_Y(Y)).$$

Ismeretes, hogy egy valószínűségi változót saját eloszlásfüggvényébe visszahelyettesítve (0,1)-en egyenletes eloszlást kapunk. Két változó rangkorrelációját a mintából megbecsülhetjük úgy is, hogy minden elemre megnézzük, hányadik a növekvő sorrendbe rendezett mintában, és ezen új változókra számulunk lineáris korrelációt.

Ezáltal elérhető az, hogy nem csak lineáris, hanem már szigorú monoton kapcsolatnál is 1 legyen a rangkorreláció. Tehát a változókat akármilyen szigorúan mondoton függvénnyel transzformáljuk, a rangkorreláció értéke nem változik (kivéve ha az egyik változót szigorúan monoton növő, a másikat csökkenő függvénnyel transzformáljuk, mert akkor az érték nem változik, de az előjel ellentétes lesz).

Egy harmadik féle korrelációs együttható a Kendall τ, amely a (2.6.)-os részben kerül tárgyalásra. Elöljáróban annyit, hogy -1 és 1 között van, és az ordinális asszociációt méri két változó között.

Mindhárom mutató problémája, hogy az összefüggést egyetlen számmal próbálják jellemezni, ám a valóságban az összefüggés jóval bonyolultabb, függhet például a valószínűségi változók nagyságától, és ezt már nem lehet mindössze egy számmal modellezni. Erre kínálnak megoldást a kopulák.Mivel az előzőekkel ellentétben nem egy számmal, hanem egy többváltozós függvénnyel jellemzi a kapcsolatot, sokkal széleskörűbb vizsgálatra ad lehetőséget.

Kopulákkal leírható például a széleken való összefüggés is, ugyanis egyes valószínűségi változók között az összefüggés változik a változók nagyságának alakulásával. Ugyanis ha normális körülmények között nem is nagy az összefüggés, például az autókban és az épületekben keletkezett károk között, extrém körülmények között (például hurrikán, tűzvész, háború, stb...) már erősen össze fognak függni, hiszen ugyanaz váltja ki őket.

Pontosan emiatt érdemes kopulákkal foglalkozni, amelyek használatosak például pénzügyi területen eszközök és hozamok vizsgálatához, egészségügyben páros szervek megbetegedéseinek elemzéséhez, és életbiztosítási területen házaspárok, vagy akár testvérek halálozásának modellezéséhez. Azért is nagyon népszerű ez a modell, mert marginális túlélési vagy eloszlásfüggvények ismerete esetén a köztük lévő összefüggőséget bevezethetjük kopulák segítségével.

5.2. Alapdefiníciók, tételek

Először is definiáljuk a kopulát, majd nézzük meg, milyen tétel alkalmazásával használható a gyakorlatban.

5.8. Definíció. *Kopulának nevezzük azokat a*

$$C : [0,1]^2 \to [0,1]$$

kétváltozós függvényeket, amelyekre teljesülnek az alábbi tulajdonságok:

1. $\lim\limits_{x_1 \to 0+} C(x_1, x_2) = \lim\limits_{x_2 \to 0+} C(x_1, x_2) = 0,$
2. $\lim\limits_{x_1, x_2 \to 1-} C(x_1, x_2) = 1$
3. $C(b,d) - c(a,d) - C(b,c) + C(a,c) \leq 0 \ (\forall\ 0 < a < b < 1,\ 0 < c < d < 1),$
4. *C minden változójában balról folytonos*
5. $\lim\limits_{x_1 \to 1-} C(x_1, x_2) = x_2$
6. $\lim\limits_{x_2 \to 1-} C(x_1, x_2) = x_1$

Az első négy tulajdonság ahhoz szükséges, hogy C két valószínűségi változó együttes eloszlásfüggvénye legyen, míg az ötödik és hatodik tulajdonság a peremeloszlásfüggvények előállítására vonatkozik.

Tehát $C : [0,1]^2 \to [0,1]$ függvény akkor kopula, ha létezik két olyan X_1 és X_2, a $[0,1]$ intervallumon egyenletes eloszlású valószínűségi változó, hogy

$$C(x_1, x_2) = P(X_1 < x_1, X_2 < x_2).$$

Természetesen kopulát nem csak két dimenzióban tudunk értelmezni, az előző képletben mindössze annyit kell változtatni, hogy két egyenletes eloszlás helyett n darabot veszünk, és az így adódó kopula $C_n : [0,1]^n \to [0,1]$ a

$$C_n(x_1, x_2, \ldots x_n) = P(X_1 < x_1, X_2 < x_2, \ldots X_n < x_n)$$

alakot ölti.

Ahhoz, hogy a kopulákat a gyakorlatban is alkalmazni lehessen, nagy segítséget nyújtott Abe Sklar 1959-es tétele, amely kátváltozós esetben így szól:

5.1. Tétel. (Sklar tétel) *Tetszőleges kétváltozós $F(x_1, x_2)$ együttes eloszlásfüggvény felírható a két peremeloszlás illetve a köztük lévő kapcsolatot leíró kopula segítségével, azaz*

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)),$$

ahol

$$F_1(x_1) = \lim_{x_2 \to \infty} F(x_1, x_2) \text{ és } F_2(x_2) = \lim_{x_1 \to \infty} F(x_1, x_2)$$

Ennek a tételnek a figyelembe vételével többdimenziós eloszlások modellezése matematikailag jelentősen könnyebbé válik, ugyanis elkülöníthetőek egymástól a peremeloszlás-függvények és a kapcsolatot leíró kopula, azaz modellezni is elég őket külön-külön.

A tétel megfordítása is igaz, azaz ha az $F(x_1, x_2)$ függvény felírható a két peremeloszlás függvényeként egy tetszőleges C kopula segítségével, azaz

$$F(x_1, x_2) = C(F(x_1), F(x_2))$$

alakban, akkor $F(x_1, x_2)$ együttes eloszlásfüggvény.

Ismerve az együttes (F), ilettve a peremeloszlásokat (F_1 és F_2), alkalmazva, hogy $F_1^{-1}(F_1(x_1)) = x_1$, látható, hogy

$$F(x_1, x_2) = F(F_1^{-1}(F_1(x_1)), F_2^{-1}(F_2(x_2))),$$

amelyből már minden ismert tehát $u = F_1(x_1)$ és $v = F_2(x_2)$ helyettesítéssel

$$C(u, v) = F(F_1^{-1}(u), F_2^{-1}(v)).$$

5.3. Nevezetes kopulák

Az összefüggő kockázatok modellezésére gyakran használják az Arkhimédészi kopulákat, mivel matematikailag könnyen kezelhetőek, kevés paraméterrel rendelkeznek, és felírhatóak zárt alakban.

5.9. Definíció. *Egy* $C : [0,1]^2 \to [0,1]$ *kopulát Arkhimédészi kopulának nevezünk, ha felírható*

$$C(u,v) = \Phi^{[-1]}(\Phi(u) + \Phi(v))$$

alakban, ahol $\Phi : [0,1] \to [0,\infty], \Phi(1) = 0$*, konvex és szigorúan monoton csökkenő.*

5.1. Megjegyzés. Φ*-t generálófüggvénynek szokták nevezni.*

A fenti definícióban alkalmazott $\Phi^{[-1]}$ jelölés Φ pszeudo-inverzét jelöli, azaz

$$\Phi^{[-1]}(t) = \begin{cases} \Phi^{-1}(t), & \text{ha } 0 \le t \le \Phi(0) \\ 0, & \text{ha } \Phi(0) \le t \le \infty \end{cases}$$

Ha $\Phi(0) = \infty$, akkor $\Phi^{[-1]} = \Phi^{-1}$.

Clayton kopula

A kopula alakja:

$$C(u,v) = (u^{-\alpha} + v^{-\alpha} - 1)^{-\frac{1}{\alpha}}$$

A generálófüggvény:

$$\Phi(t) = t^{-\alpha} - 1, \text{ ahol } \alpha > 0$$

Ez egy aszimmetrikus kopula, a bal széleken mutat nagyobb összefüggést, azaz ha az egyik kár kicsi, valószínűleg a másik kár is hasonlóan kicsi lesz.

Gumbel kopula

A kopula alakja:

$$C(u,v) = \exp\{-((-\ln u)^{\alpha} + (-\ln v)^{\alpha})^{\frac{1}{\alpha}}\}$$

A generálófüggvény:

$$\Phi(t) = (-\ln t)^{\alpha}, \text{ ahol } \alpha \geq 1$$

A Clayton kopulához hasonlóan ez szintén egy aszimmetrikus kopula, de a jobb széleken mutat nagyobb összefüggést, azaz ha az egyik kár nagyon nagy, a másik is valószínűleg nagy lesz, ennek oka lehet például valamilyen természeti katasztófa, tűzeset.

Frank kopula

A kopula alakja:

$$C(u,v) = -\frac{1}{\alpha}\ln\left(1 + \frac{(e^{-\alpha u}-1)(e^{-\alpha v}-1)}{e^{-\alpha}-1}\right)$$

A generálófüggvény:

$$\Phi(t) = -\ln\left(\frac{e^{\alpha t}-1}{e^{\alpha}-1}\right), \text{ ahol } \alpha \neq 0$$

A Frank kopula az előző kettővel ellentétben szimmetrikus.

További kopulákról a [4] könyvben lehet olvasni, én csak a három leggyakrabban használtról írtam.

6. fejezet

Hasonlóságok és különbségek

Az előző fejezetben tárgyalt kopulák, és a negyedik fejezetben bemutatott frailty modellek is alkalmazhatóak kétváltozós túlélési függvények modellezéséhez. Ám a két módszer más irányból közelíti meg a kérdést (lásd [1]).

Az összehasonlításhoz [6] alapján tekintsük a kutyáknál a csípőtörés gyógyulásáig eltelt időt. A diagnózis megállapításához két módszert használnak, röntgent illetve ultrahangot. Az ezen diagnózisok által kapott gyógyulási idők lesznek a túlélési idők. Természetesen a két túlélési idő között erős összefüggés van.

Erről azonban a marginális Cox regressziós modellek nem adnak információt. Az összefüggés modellezéséhez használhatóak kopula modellek, illetve frailty modellek. De mi az eltérés a két megközelítés között?

A kopula modelleknél a két túlélési idő együttes túlélésfüggvénye, azaz a kopula a két marginális túlélésfüggvény függvényeként van modellezve, és ez határozza meg a függés típusát.

Általában két lépésből áll a modellezés, az első lépésben a marginális túlélésfüggvényeket becsüljük, paraméteresen, szemiparaméteresen vagy nemparaméteresen. Ezután, a második lépésben a kopula függvény paramétereinek becslése következik, maximum likelihood módszerrel, úgy hogy a likelihoodban kicseréljük a marginális túlélésfüggvényeket az első lépésben becsült változattal.

Ezzel szemben a frailty modell egy feltételes hazárd modell, ahol létezik egy véletlenszerű hatás, azaz a frailty, amely multiplikatívan hat a hazárdrátára. Továbbá a feltételes túlélési idők függetlenek, ha a feltételben a frailty szerepel.

Ebben az esetben az együttes sűrűségfüggvényt úgy kapjuk meg, hogy kiintegráljuk a frailty-t a kétváltozós feltételes túlélés eloszlásból. Belátható, hogy az együttes sűrűségfüggvény egy Archimédeszi kopula alakját fogja felvenni. Azaz a frailty modell megfelel egy adott Archimédeszi kopulának, annak ellenére, hogy a két modell más oldalról közelíti meg a problémát.

7. fejezet

Adatok elemzése

7.1. Az adatokról általában

A modellezéshez rendelkezésre álló adataimat Tárnok Edinától kaptam, akinek Tusnády Paula bocsátotta a rendelkezésére. Az adatok egy temetőből származnak, olyan párokról akiknek a neve alapján feltehető volt, hogy házasok voltak.

Az adathalmaz 483 párról tartalmazza a férj illetve a feleség születési és halálozási évét. Ezen adatok közül a legkorábbi 1823-as, a legkésőbbi pedig 2005-ös. A vizsgált párok közötti átlagos korkülönbség valamivel több, mint 5 év. A férfiak átlagéletkora körülbelül 71, még a nőké 74 év. A 483 pár közül 376 esetben a férfi, 74 esetben pedig a nő volt az idősebb, 33 alkalommal ugyanabban az évben születtek. Ezek a számok egybeesnek azzal az általános képpel, hogy a nők általában tovább élnek, mint a férfiak, illetve többségében ők a fiatalabbak egy házasságban.

Sajnos a 483-as elemszám kevés ahhoz, hogy messzemenő következtetéseket vonhassunk le az eredményekből, ám alkalmas lehet arra, hogy néhány általános jelenséget megfigyelhessünk rajta. További probléma az adatokkal, hogy széles skálán mozognak, így torzulhatnak az eredmények, hiszen a születéskor várható élettartam nőtt az eltelt idő folyamán.

Ha az évek mellett a halál pontos napja is rendelkezésünkre állna, további elemzéseket

készíthetnénk arra vonatkozólag, hogy mennyire rövid időn belül történtek egymáshoz képest az elhalálozások.

A házaspár tagjainak élettartamának egymáshoz való viszonyát vizsgálhatjuk többek között kopulákkal és túlélési modellekkel is.

7.2. Kopulák

Tárnok Edina a párok élettartamának modellezéséhez a 4. fejezetben bemutatott kopulákat használta (lásd [10]). Számításai alapján ezen adathalmazban az élettartamok összefüggőségét a legjobban az 5. fejezetben bemutatott Clayton-kopula írta le.

Ezután kiszámolta a két-, illetve több életre szóló biztosítások díját először függetlenséget feltételezve, majd a becsült Clayton kopula alapján. Eredményül jelentős eltéréseket kapott a kétféle módon kiszámolt díjakra, a kockázati biztosítások ugyanis felülkalkuláltak, az elérési biztosítások pedig alulkalkuláltak független élettartam adatokkal számolva.

Emiatt tehát egyértelműen elmondható, hogy hatással van a házaspár tagjainak élettartama egymásra, és érdemes ezzel a témával mélyebben foglalkozni. Én ezt a következőben túlélési modellek segítségével próbálom belátni.

7.3. Kaplan-Meier modell

Az egyik leggyakrabban használt túlélési modell a Kaplan-Meier modell. Először ezt fogom alkalmazni a rendelkezésemre álló adathalmazon. A vizsgálat során a célom annak feltárása, hogy a pár egyik tagjának halála hat-e a másik tag túlélési valószínűségére.

Ilyen modellek futtatására számos lehetőségünk nyílik, hiszen alkalmas rá többek között az SPSS, Matlab, SAS vagy az R programnyelv. Én az adatok könnyebb kezelhetősége miatt R-ben programoztam. A *survival* csomagot használtam, és a [5]-ben található módszertant követtem. Ebben az alfejeztben a Kaplan-Meier modellt fogom

alkalmazni, amely az egyik leggyakrabban használt túlélési modell.

Az elemzéshez létrehoztam több idő és státuszváltozót is, függően attól, hogy melyik fázisban - a pár halála előtt vagy után - vagyunk-e.

7.3.1. Első fázis

Az első vizsgált időtartam addig az időpontig tart, amíg valaki meg nem hal a házaspár tagjai közül. Ilyenkor az illető halálára nem hat a párja elvesztése.

A módszertan ismertetéséhez vezessük be az alábbi jelöléseket egy adott házaspár tagjaira:

- D_f: A házaspár férfi tagjának halálának időpontja
- D_n: A házaspár női tagjának halálának időpontja
- b_f: A házaspár férfi tagjának születésének időpontja
- b_n: A házaspár női tagjának születésének időpontja
- t_f: A házaspár férfi tagjához tartozó időváltozó
- t_n: A házaspár női tagjához tartozó időváltozó

Az első fázisnak akkor van vége, mikor a pár valamelyik tagja meghal. A fázis végét jelöljük S_f-fel, illetve S_n-nel egy adott párra. Tehát az első fázis vége

$$S_f = S_n = \min(D_f, D_n).$$

Az időváltozók az eddig megélt évek számát mutatják, azaz az első fázisban

- $t_f = S_f - b_f$,
- $t_n = S_n - b_n$.

Ez a változó a pár egyik tagjára megegyezik az élettartammal is. Az általam vizsgált adatfájlban 11-szer fordul elő, hogy a pár mindkét tagja ugyanabban az évben halt meg, ebben az esetben mindkét félnél megegyezik az élettartammal.

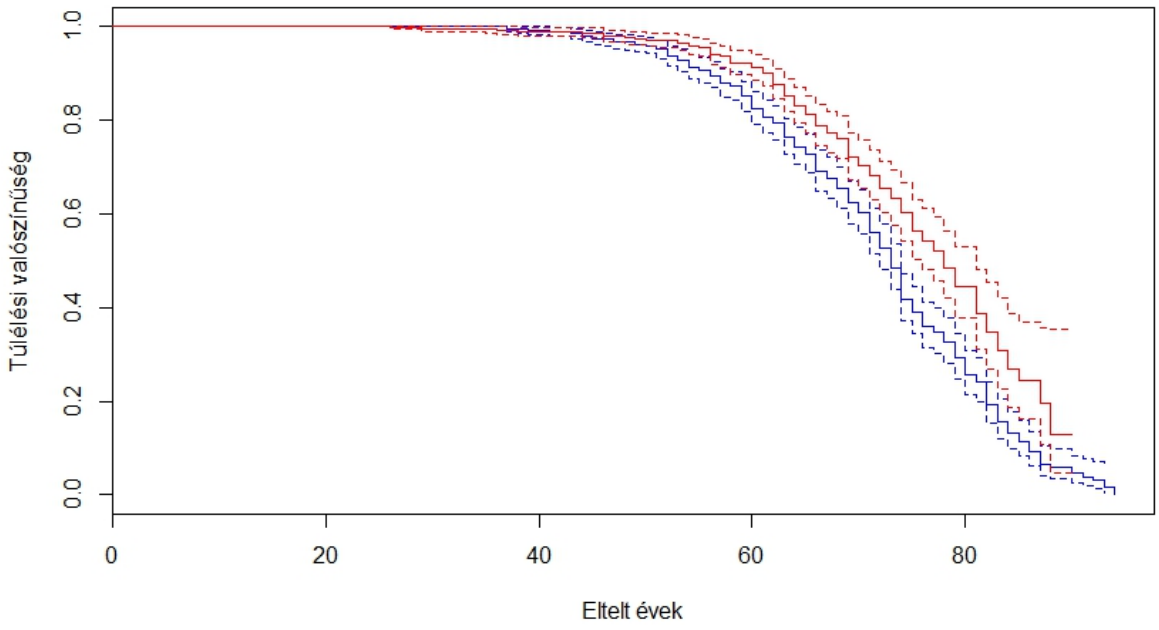
A státuszváltozó abban az esetben lesz 1, ha valaki meghal a vizsgált fázisban, azaz a férfiak státuszváltozója

$$e_f = \begin{cases} 0, & \text{ha } S_f < D_f \\ 1, & \text{különben.} \end{cases} \tag{7.3.1}$$

A nők státuszváltozója hasonlóképpen

$$e_n = \begin{cases} 0, & \text{ha } S_n < D_n \\ 1, & \text{különben.} \end{cases} \tag{7.3.2}$$

Először ezekre a státusz-, illetve időváltozókra futtattam le a Kaplan-Meier modellt. Az alábbi ábrán a nem, mint magyarázóváltozó szerinti bontás látható.



7.1. ábra. Kaplan-Meier modell az első fázisban

A (7.1) ábráról megfigyelhető, hogy a nők túlélési görbéi minden időpontban magasabban helyezkednek el, mint a férfiakéi. Ezt alátámaszthatjuk χ^2 statisztikával is, amelynek nullhipotézise az, hogy két túlélési görbe között nincs szignifikáns különbség. Lefuttatva a tesztet a p-érték 0,00000241, ezért elutasítható az a nullhipotézis. Tehát a statisztika és az ábra is alátámasztja azt a feltevésünket, hogy a nem fontos prediktor a túlélési valószínűségek tekintetében, legalábbis az első fázisban.

A gyakran használt, ugyanezzel a nullhipotézissel rendelkező log-rank tesztet alkalmazva hasonló nagyságrendű p-értéket kapunk, ezzel is bizonyítva azt, hogy van különbség a férfiak és a nők túlélési görbéi között ebben a fázisban. A különbség a tesztek között mindössze a kilépési idők súlyozása a statisztika kiszámolásánál, ami általában hasonló eredményre vezet.

Látható továbbá, hogy az esemény (tehát a halál) 344 alkalommal férfiaknál következett be, 150 alkalommal pedig nőknél. Ez összesen 494 esemény, mivel 11 alkalommal ugyanabban az évben halt meg a pár.

7.3.2. Második fázis

Ezután röviddel a pár halála utáni időszakot vizsgáltam, hiszen több tanulmány szerint is kimutatható, hogy gyakran egy-két évvel a férj/feleség halála után a hirtelen sokk és változás miatt a pár másik tagja is meghal (úgynevezett "broken-heart syndrome"). Tehát ez az időszak kritikus lehet a társát túlélő partner számára.

Először megnéztem a túlélési görbék alakulását egy évvel a házastárs halála után következő időszakban. Tehát ebben a fázisban csak azokat az embereket vizsgálom, akik már elvesztették a párjukat, de ők még élnek. Emiatt $483 - 11 = 472$ alany túlélési valószínűségeit fogom vizsgálni. Természetesen, ha a halálozási év mellett az adatfájl tartalmazná a pontos dátumot, akkor még rövidebb intervallumot is vizsgálhatnánk.

A második fázis vége azon férfiakra, akik túlélték a feleségüket:

$$S_f = \min(D_f, D_n + 1)$$

És azon nőkre, akik túlélték a férjüket:

$$S_n = \min(D_n, D_f + 1)$$

Az esemény akkor következik be, ha a vizsgált egyén meghal az időszakban és akkor lesz cenzorált a megfigyelés, ha a minimum második tagja lesz a kisebb, azaz előbb jár le az egy év a társ halála után, minthogy megtörtént volna az esemény.

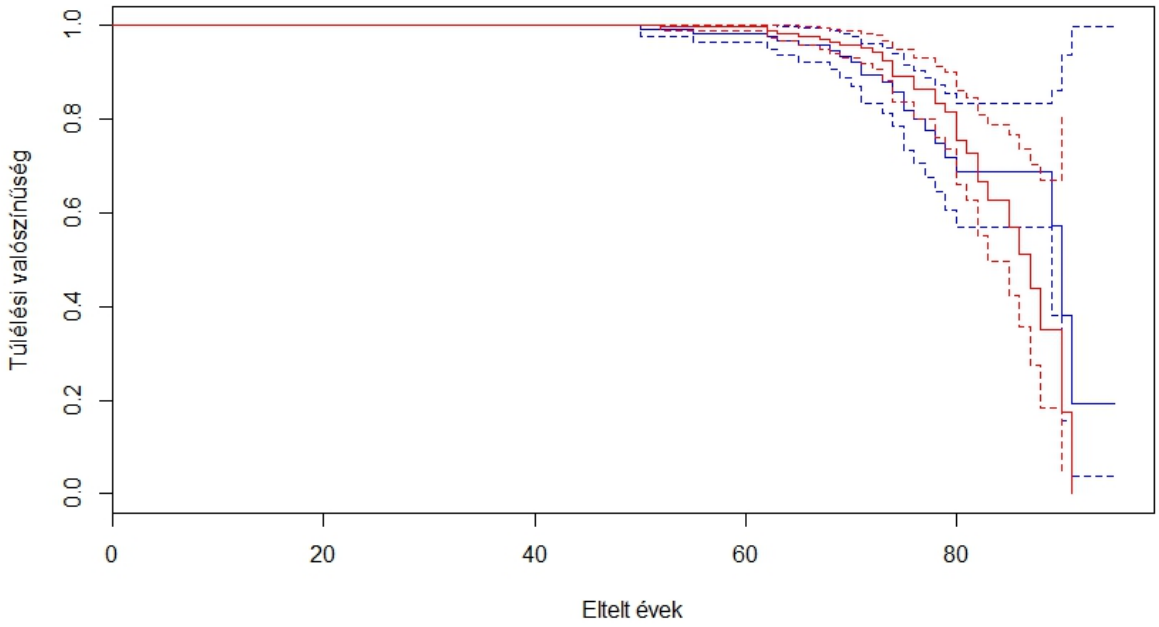
Az időváltozók továbbra is a fázis vége és a születés közti időtartamot jelölik, míg a státuszváltozó is abban az esetben lesz 1, ha az illető elhalálozik a vizsgált fázis vége előtt, azaz például a férfiakra:

$$t_f = S_f - b_f,$$

$$e_f = \begin{cases} 0, & \text{ha } S_f < D_f \\ 1, & \text{különben.} \end{cases} \tag{7.3.3}$$

Ezen változókra lefuttatva a Kaplan-Meier modellt, azt az eredményt kapjuk, hogy a férfiak esetében mindössze 10-szer, a nők esetében pedig 19-szer következett be a vizsgált esemény. Mivel ezek a számok annyira kicsik, hogy további vizsgálatokra alkalmatlanok, így megpróbálkoztam a házastárs halála utáni két éves intervallumot vizsgálni. Ilyenkor mindössze a második fázis végét jelző S_f és S_n fog módosulni, hiszen a második tagban egy helyett kettő fog állni, minden más marad ugyanúgy.

Ebben az esetben a férfiak esetében 22-szer, a nőknél pedig 33-szor következett be halál kevesebb, mint két évvel a pár halálát követően. A túlélési görbék alakulását a (7.2) ábrán láthajuk. Körülbelül 80 éves korig a nők piros túlélési görbéje a férfiaké felett halad, ám ott - feltehetően a kevés elemszám és néhány kiugró érték következtében - a kék görbében egy ugrás nagy található. Mivel a nők között többször következett be halál a vizsgált fázisban így az ő görbéjük simább. Alkalmazva a χ^2 és a log-rank tesztet azt kapjuk eredményül, hogy a két túlélésgörbe között nincs szignifikáns különbség.



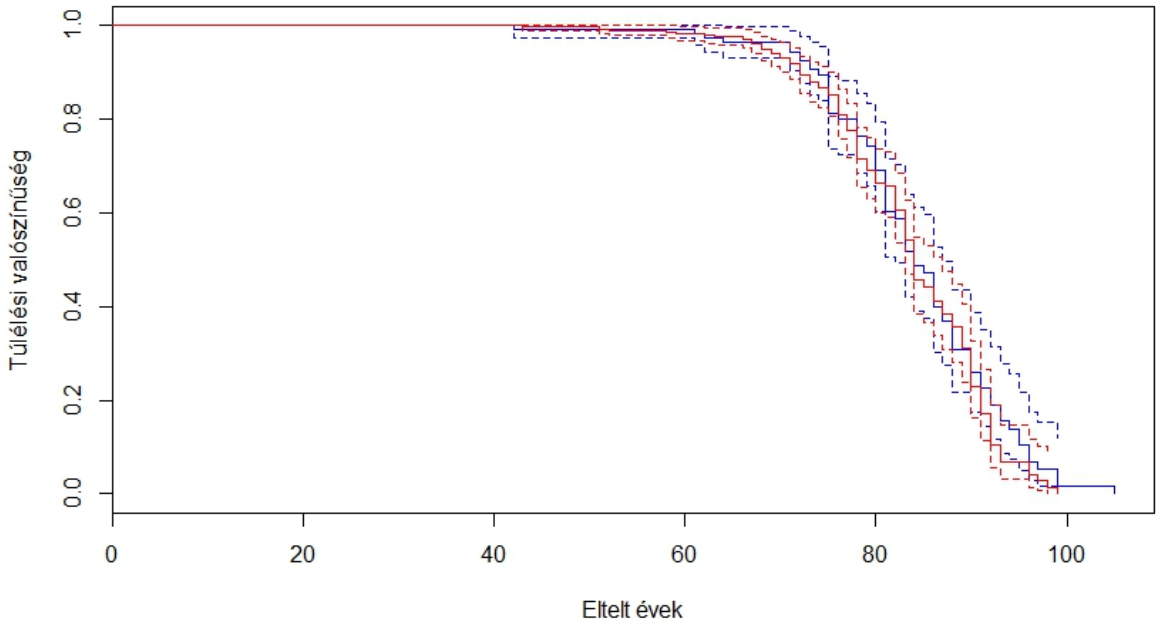
7.2. ábra. Kaplan-Meier modell a második fázisban

7.3.3. Harmadik fázis

Most pedig vizsgáljuk a hosszútávú következményeit a házastárs elvesztésének. Itt már csak azokat az embereket fogjuk vizsgálni akik túlélték a párjukat (tehát az első fázist), és az azt követő két évet (azaz a második fázist).

Ezekre az emberekre a harmadik fázis vége megegyezik a halál időpontjával, azaz a férfiakra $S_f = D_f$ és a nőkre $S_n = D_n$. A mi adathalmazunk olyan emberek születési és halálozási éveiből áll, akik már meghaltak, azaz nincsenek cenzorált elemeink. Emiatt a státuszváltozó most mindenhol 1 lesz. Az időváltozók pedig az előbbiekhez hasonlóan a születés és a harmadik fázis vége között eltelt időt jelölik, tehát $t_f = S_f - b_f$ és $t_n = S_n - b_n$, amely most megegyezik az életkorral.

Lefuttatva a Kaplan-Meier modellt, a túlélési görbék a (7.3) ábrán láthatóak.



7.3. ábra. Kaplan-Meier modell a harmadik fázisban

Első ránézésre is észrevehető, hogy a két túlélési görbe egymáshoz nagyon közel halad, nincs akkora eltérés köztük, mint például az első fázisban. Alkalmazva a χ^2 és a log-rank teszteket, 0,4 körüli p-értékekeket kapunk, azaz nem utasítható el a két túlélésgörbe azonosságáról szóló nullhipotézis, azaz az ábrán látható sejtésünk beigazolódott. Tehát ez az eredmény arra enged következtetni - bár a mintám mérete túl kevés ahhoz, hogy ezt teljes bizonyossággal állíthassuk -, hogy a túlélés valószínűségére a harmadik fázisban már nem hat megkülönböztetőleg a nem.

7.3.4. További elemzések

A házaspárok élettartamának egymásra való hatását további magyarázóváltozókkal is vizsgáltam. A pár tovább élő tagjának túlélését elemeztem az előbbiekhez hasonlóan.

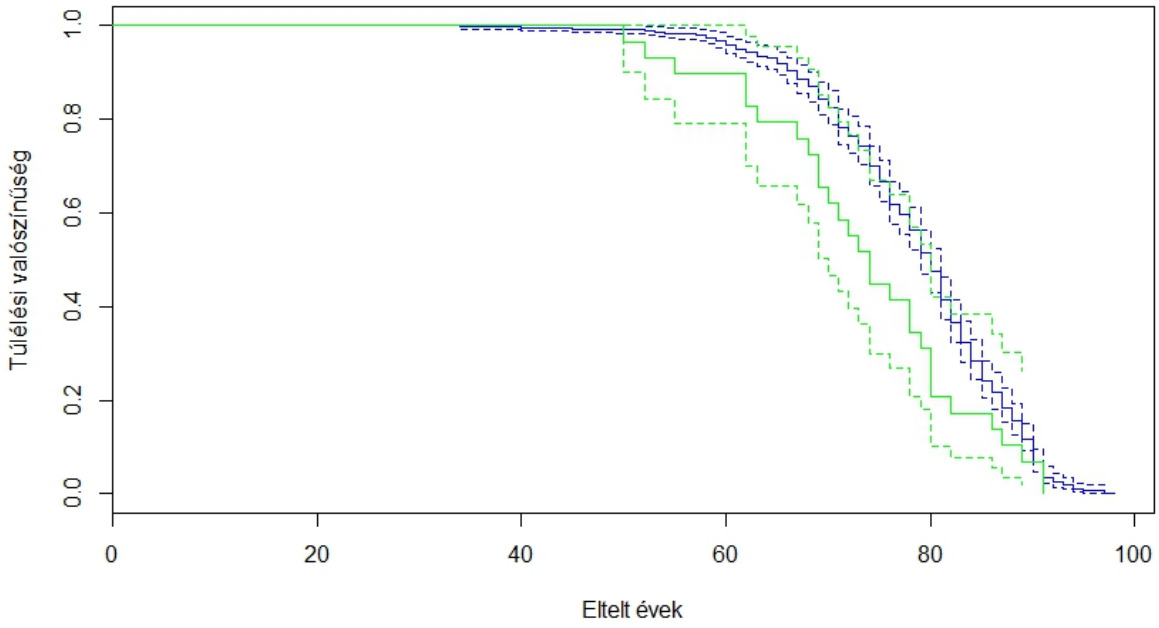
Magyarázóváltozónak létrehoztam egy új dummy változót, a második fázishoz hasonlóan csak azokra az emberekre, akik túlélték a párjukat. A változó akkor vett fel 1-et ha

a házaspár tagjainak halála között eltelt idő legfeljebb egy év volt (természetesen mivel nem ismerjük a halál pontos napját, így ez akár majdnem két év lehet).

Azaz például a férfiakra

$$m_f = \begin{cases} 1, & \text{ha } |D_f - D_n| \leq 1 \\ 0, & \text{különben.} \end{cases} \tag{7.3.4}$$

Ezt felhasználva magyarázóváltozóként a Kaplan-Meier modellben az alábbi túlélési görbéket kapjuk:

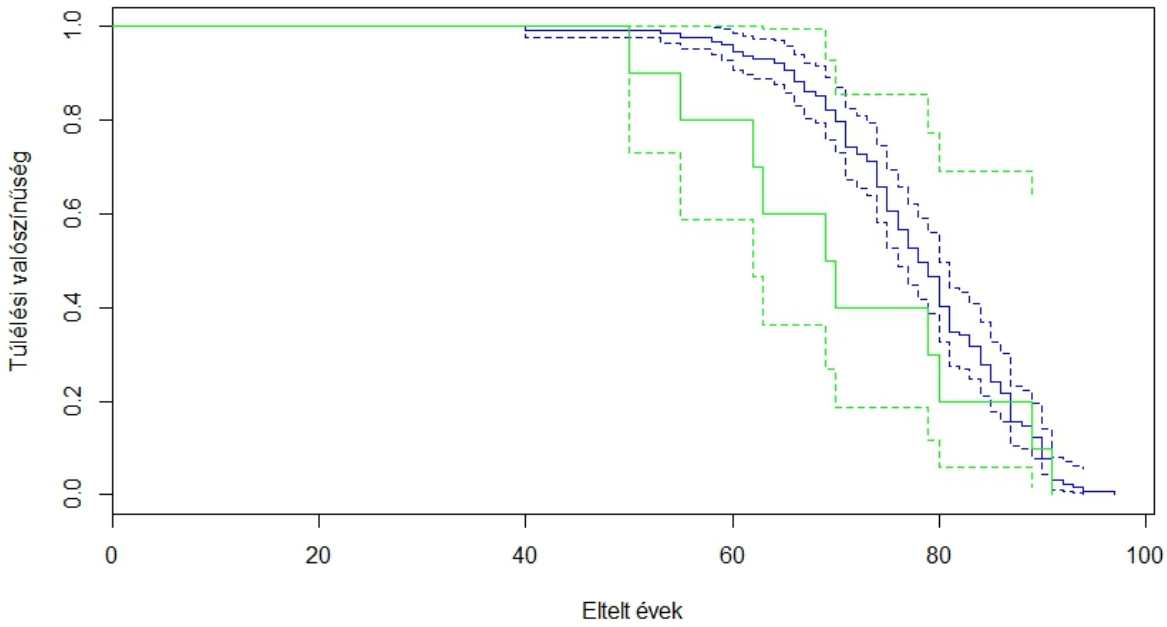


7.4. ábra. Kaplan-Meier modell egyéb elemzéshez

A (7.4) ábrán kék görbe az m magyarázó változó 0, a zöld pedig az 1 értékéhez tartozik. Az $m = 0$ eset azt jelenti, hogy a túlélő tag már több, mint egy évvel túlélte társa halálát. A zöld görbe pedig azon esetekhez tartozik, mikor még friss a házastárs halála. Szemmel láthatóan a zöld görbe a kék alatt halad, bár a hozzá tartozó konfidencia intervallum felső határa belemetsz a kék görbébe.

A χ^2 teszt alapján a p-érték 0,0179, azaz 5%-os szignifikanciaszinten elutasíthatjuk a nullhipotézist, tehát van jelentős különbség a két túlélési görbe között. Szóval az ezen a mintán elvégzett számítások alapján elmondhatjuk hogy a házastárs halála - legalábbis rövidtávon - hatással van az életben maradt túlélési valószínűségére.

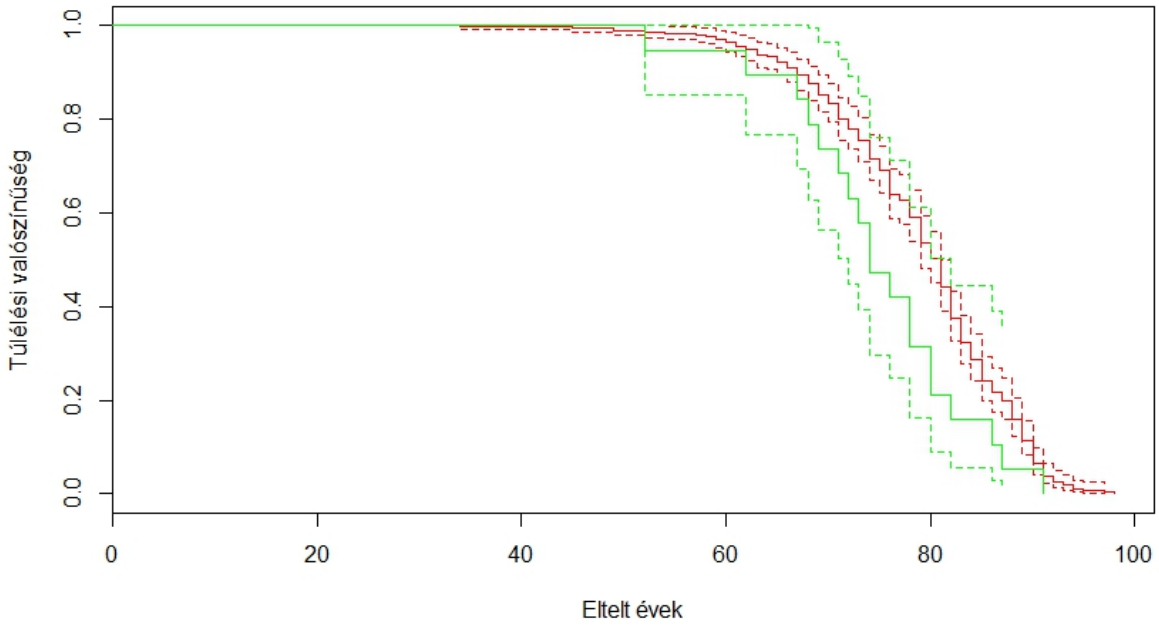
De vajon ez a hatás ugyanolyan mértékű a nőkre és a férfiakra is? A kérdés megválaszolásához futtassuk le ezt az analízist külön-külön mindkét nemre. Vegyük először a férfiakat. A (7.5) ábrán a zöld vonal jelzi az előzőekhez hasonlóan azon férfiak túlélési valószínűségeit, akik nemrég vesztették el a párjukat.



7.5. ábra. Kaplan-Meier modell egyéb elemzéshez - férfiak

Habár az ábrán a zöld görbe egyértelműen a kék alatt halad, látható, hogy nem olyan sima mint az, hiszen csak 10 ilyen halál történt. A kevés elemszám miatt nem érdemes messzemenő következtetéseket levonni, amit a χ^2 statisztika is megerősít. A kapott p érték 0,278, ami alapján nem vethető el az a nullhipotézis, hogy nincs különbség a két csoport túlélési valószínűségei között.

Tekintsük most a nők túlélési görbéit (6. ábra).



7.6. ábra. Kaplan-Meier modell egyéb elemzéshez - nők

Itt már a χ^2 statisztika is alátámasztja 0,0261-es p-értékkel, hogy a párjukat nem rég elvesztő nők rosszabb túlélési esélyekkel rendelkeznek, mint a többiek. Habár ebben az esetben itt is csak 19 esetben történt halál, ezért biztos következtetésekre nem alkalmasak az eredmények, de a most kapott értékek összhangban állnak a harmadik fázis eredményeivel. Mivel a nők várható élettartama alapesetben hosszabb, mint a férfiaké, de mégis az látható hogy szignifikánsan nem különbözik ez az életkor a harmadik fázisban, ezáltal arra következtethetünk, hogy a házastárs elvesztése a nők túlélési esélyeit jobban rontja, mint a férfiakéit. Mint ahogy ezt a most megkapott p-értékek is alátámasztják. Természetesen hangsúlyozandó, hogy az elemzéshez használt adatfájl mérete még az 1000 embert sem haladja meg, így a következtetések nem feltétlenül állják meg a helyüket általánosságban.

A második fázisban leírtak alapján ismét kipróbáltam egy módosított dummy válto-

zóval is az eljárást, amikor két évre növeltem a pár tagjainak halála között eltelt időt, azaz például a férfiakra legyen most minden párra

$$m_f = \begin{cases} 1, & \text{ha } |D_f - D_n| \leq 2 \\ 0, & \text{különben.} \end{cases} \tag{7.3.5}$$

Ebben az esetben a görbék hasonlítanak a 4., 5. és a 6. ábrán látottakhoz, mindössze a nagyobb intervallum miatt a zöld görbék is simábbak. A χ^2 statisztikák alapján az együttes esetben a p-érték 0,00128, tehát van megkülönböztető ereje m-nek. Most az előzővel ellentétben a férfiakra és a nőkre is 0,02 körül alakul a p-érték, tehát egyik esetben sem tartható meg a túlélési valószínűség egyezőségére vonatkozó nullhipotézis 5%-os szignifikanciaszinten.

7.4. Cox regresszió

Ahhoz, hogy több faktor együttes hatását vizsgálhassuk az élettartamokra, futtassunk most Cox regressziót. Hasonlóan a Kaplan-Meier modellhez, most is bontsuk három részre az élettartamokat.

7.4.1. Első fázis

Mivel rendelkezésre állnak az életkorok mellett a születési dátumok is, készítettem egy új magyarázóváltozót, amely azt jelzi, hogy az illető melyik évtizedben született. Mivel a várható élettartam folyamatosan nő, ennek lehet befolyásoló hatása az eredményre. Az első elemzésbe tehát ezt a változót, illetve a nemet vettem be, mint magyarázó változókat. Lefuttatva a regressziót az alábbi eredményeket kapjuk:

	coef	exp(coef)	se(coef)	z	Pr(> |z|)
nem	-0,52	0,60	0,10	-5,21	0
születési évtized	0,20	1,22	0,04	5,36	0

7.1. táblázat. Cox regresszió az első fázisban

Az utolsó oszlopban lévő számok az egyes magyarázóváltozók megkülönböztető erejére vonatkoznak. Akkor szignifikáns egy változó, ha ez az érték kevesebb, mint 0,05. Mivel ez mindkét sorra teljesül, ezért elmondhatjuk hogy egy adott ember túlélési valószínűségére szignifikáns hatással van a neme és a születési évtizede. (A táblázatban szereplő értékeket kerekítettem, az utolsó oszlopban például nem nullák, hanem nagyon kicsi számok szerepeltek.)

Most nézzük az első oszlopban található együtthatókat. A pozitív előjel azt jelenti hogy a változó magasabb értékeire a halál esélye is nagyobb. Az alany nemére vonatkozóan ez az együttható negatív, tehát a magasabb értékű változó túlélési esélyei jobbak, ami a mi kódolásunkban a női alanyokra vonatkozik. Ez a megfigyelés egybeesik az eddig tapasztaltakkal. Mivel a Cox regresszió egy arányos hazárd modell, ezért ha egy adott megfigyelésnél a másik magyarázóváltozót nem, csak a nemet változtatom meg férfiról nőre, az 0,6-szorosára csökkenti a hazárdrátát, tehát 40%-kal.

A születési évtized változóra ez az érték pozitív, ami azt a meglepő eredményt adja, hogy minél későbbi évtizedben született valaki, annál rosszabbak a túlélési esélyei.

Első pillantásra azért nem tűnik logikusnak ez az eredmény, mert a vizsgált időszakban, tehát a 19. század végén és 20. század elején a születéskor várható élettartam jelentősen megnőtt, és azt várnánk, hogy a mi elemzésünk is ezt fogja tükrözni. Ám ha a születés helyett a halál évét rakjuk be magyarázóváltozónak, ott már azt az eredményt kapjuk, hogy minél későbbi évben halt meg valaki, annál jobbak voltak a túlélési esélyei.

Ennek a különös eredménynek az oka lehet például a kevés, nem reprezentatív mintaelemszám. Továbbá az adatok egy temetőből származnak, így nagyon régi születésűek csak akkor kerülhettek be az adatfájlba, ha sokáig éltek, mivel például a régebbi sírokra

gyakran rátemetkeznek.

További statisztikákat is vizsgáltam, például a modellhez tartozó Cox & Snell R^2 értéke mindössze $5,3\%$, ami azt jelenti, hogy mindössze ezekkel a magyarázóváltozókkal csak nagyon gyenge becslést adhatunk az élettartamra, ami érthető is, hiszen a nem és a születési évtized mellett számos tényező befolyásolja életünk hosszát, többek között az egészségügyi állapot, dohányzás, alkoholfogyasztás, örökletes betegségek és természetesen a véletlen balesetek is. Ettől függetlenül a p-értékekből látható, hogy ezen változóknak is szignifikáns magyarázóereje van.

Az output tartalmaz még három statisztikai tesztet is (likelihood arány teszt, Wald teszt, Score teszt), amelyekkel tovább vizsgálhatjuk a modell jelentőségét. Mindhárom esetben a nullhipotézis az, hogy az együtthatók mind nullák, azaz a kiválasztott változóknak nincs magyarázóereje. Mindegyik teszt alapján elutasítható ez a nullhipotézis, hiszen a p-értékek ismét nagyon közel vannak a nullához. A három eljárás egyébként nagy minta esetében hasonló eredményt ad, kis mintánál, mint jelent esetben, a likelihood arány teszt a leggyakrabban használt.

7.4.2. Második fázis

Futtassuk most a Cox regressziót azon emberekre, akik a közelmúltban – nem több, mint két éve – vesztették el a párjukat. Az alábbi eredményeket kaptuk:

	coef	exp(coef)	se(coef)	z	Pr(> |z|)
nem	-0,13	0,87	0,29	-0,46	0,64
születési évtized	0,20	1,21	0,12	1,61	0,11

7.2. táblázat. Cox regresszió a második fázisban

Az utolsó oszlopból leolvasható p-értékek szerint a vizsgált változóknak ebben a fázisban nincs megkülönböztető ereje. Ezzel összhangban áll a likelihood arány, a Wald és a Score teszt 0,27-es p-értéke is, hiszen emiatt nem tudtuk elutasítani a nullhipoté-

zist, hogy minden együttható nulla a modellben. Az R^2 értéke is $0,06\%$-ra csökkent. A Kaplan-Meier modellnél is hasonló eredményeket kaptunk ebben a fázisban.

7.4.3. Harmadik fázis

Végül nézzük, hosszú távon milyen eredményeket kapunk.

	coef	exp(coef)	se(coef)	z	Pr(> |z|)
nem	-0,02	0,98	0,15	-1,11	0,915
születési évtized	0,33	1,39	0,06	5,33	0

7.3. táblázat. Cox regresszió a harmadik fázisban

Az utolsó oszlopban található p-értékeket megfigyelve látható, hogy a nemnek nincs magyarázóereje a modellben a harmadik fázisban, ahogy ezt már megfigyeltük a Kaplan-Meier modellnél is. Tehát ugyanarra az eredményre jutottunk mint korábban, vagyis a nem a harmadik fázisban már nem meghatározó tényező a túlélési valószínűségeket tekintve.

A születési évtized még mindig fontos prediktor lehet, de ennek valószínűleg az adatok érdekes eloszlásához van köze. Az átlagéletkor például nagyot csökken az 1920-as években születetteknél, de ezzel együtt a mintaelemszám is kevesebb, mint a felére esett vissza. Ezeknek az eredményeknek az oka szintén az adatok speciális beszerzési helye lehet. Ugyanis egy temetőben ahogy haladunk előre az időben, egyre kevesebb adott évben született mintánk lesz, és a kevés mintaelemszám miatt esetleges kiugró értékek jobban torzíthatják a modellünket.

7.4.4. További elemzések

Lefuttatva a Cox regressziót a (7.3.4)-es alfejezetben leírt változókra, magyarázóváltozóként felhasználva a nemet, születési évtizedet és a létrehozott m változót, az így létrejött

modellben a születési évtized és m lesznek szignifikáns magyarázóváltozók, a modell Cox & Snell R^2-ének az értéke pedig 11,3%-ra nő.

8. fejezet

Összefoglalás

A szakdolgozat célja elsősorban különböző valószínűségi változók közötti összefüggés modellezése. A dolgozat első felében megismerkedtünk a két leggyakrabban használt módszer, a túlélési modellek és a kopulák matematikájával. Szó esett a gyakran használt frailty modellekről is, összevetve a hasonlóságokat és különbségeket a korábbiakkal.

Ezután egy speciális adathalmazon futattam a először Kaplan-Meier modellt, majd a Cox regressziót. Az adathalmaz házaspárok születési, és halálozási évét tartalmazta. Az elemzéseket három részre bontottam, hogy vizsgálhassam a házastárs elvesztésének halandóságra gyakorolt hatását rövid és hosszú távon is.

A vizsgálatok során érdekes eredményeket tapasztaltam. Az első fázisban, amikor még mindkét fél életben van, az alany neme fontos prediktor a modellekben. Ez előzetesen is vártuk, hiszen köztudomású és számos statisztika által alátámasztott tény, hogy a nők tovább élnek, mint a férfiak. A második fázisban, mikor is a házastárs halála utáni rövid intervallumot vizsgálom, már nincs szignifikáns eltérés a férfiak és a nők túlélési görbéje között. A harmadik fázisba már jóval több megfigyelés esik, mint a másodikba, és itt is arra a meglepő megállapításra jutottunk, hogy ebben az időszakban már nem hat szignifikánsan az alany neme a túlélési valószínűségre.

Emellett elemeztem a túlélési valószínűségeket két csoportra bontva a populációt: akik egy-két éven belül vesztették el a párjukat, illetve akik több ideje. Statisztikai

tesztek bizonyították, hogy azon emberek túlélési valószínűségei rosszabbak, akiknek a közelmúltban kellett megbirkózni a társuk halálával.

A modellt természetesen számos módon lehetne továbbfejleszteni. Az adathalmazom elemszáma sajnos kevés ahhoz, hogy a következtetéseink biztosan helytálljanak, így ugyanezeket az elemzéseket egy jóval nagyobb mintára lefuttatva további érdekes eredményeket kaphatunk. Vizsgálható lenne egyes időszakok, például a világháborúk hatása a halandósági valószínűségre, akár külön megvizsgálva a férfiakat és nőket.

Továbbá én mindössze a születési és halálozási éveket ismertem a párokról, és a Cox regresszióban például számos más magyarázóváltozó is javíthatja az előrejelzést, például vagyoni helyzet, egészségügyi állapot vagy káros szokások. Illetve pontosabb elemzéseket készíthetnénk, főleg a második fázisról, ha az év mellett a születés és halál hónapja is napja is rendeklezésünkre állna.

Végezetül tehát elmondhatjuk, hogy a házastárs halála nagyon fontos tényező egy ember túlélési valószínűségét tekintve, és a biztosító társaságoknak mindenképp érdemes lenne számolni ezzel a tényezővel is rövid-, illetve hosszútávon, mikor halandósági valószínűségeket kalkulálnak.

Irodalomjegyzék

[1] Goethalsa, K.; Janssenb, P.; Duchateaua, L.: *Frailty models and copulas: similarities and differences*, Journal of Applied Statistics, 2008

[2] Hougaard, Philip (1995): *Frailty Models for Survival Data*, Kluwer Academic Publishers

[3] Kleinbaum, David G.; Klein, Mitchel (2012): *Survival analysis: A Self-learning text*, Springer

[4] Panjer, Harry H. (2006): *Operational Risk: Modeling Analytic*, John Wiley & Sons

[5] Rétallér Orsolya (2017): *What doesn't break you makes you stronger?*, Szakdolgozat - Budapesti Corvinus Egyetem

[6] Risselada, M., van Bree, H.; Kramer, M.; Chiers, K.; Duchateau, L.; Verleyen, P.; Saunders, J. H. (2006): *Evaluation of nonunion fractures in dogs by use of B-mode ultrasonography, power Doppler ultrasonography, radiography and histologic examination*, John Wiley & Sons

[7] Rodrígez, Germán (2001): *Multivariate Survival Models*, Princeton University

URL: `http://data.princeton.edu/pop509/MultivariateSurvival.pdf`

[Letöltve: 2018.05.09]

[8] Rodrígez, Germán (2001): *Unobserved Heterogeinity*, Princeton University

URL: http://data.princeton.edu/pop509/UnobservedHeterogeneity.pdf

[Letöltve: 2018.05.09]

[9] Shemyakin, A. E.; Younz, Heekyung (2006): *Copula models of joint last survivor analysis*, John Wiley & Sons

[10] Tárnok Edina (2011): *A férj és feleség élettartamának modellezése több életre szóló életbiztosítási szerződéseknél*, Szakdolgozat - Eötvös Loránd Tudományegyetem

URL: http://web.cs.elte.hu/blobs/diplomamunkak/msc_actfinmat/2011/tarnok_edina.pdf

[Letöltve: 2018.05.09]

[11] Vékás Péter (2012): *Összefüggő biztosítási kockázatok modellezése*, Budapesti Corvinus Egyetem

URL: http://unipub.lib.uni-corvinus.hu/2093/1/VekasPeter_Osszefuggo_wp.pdf

[Letöltve: 2018.05.09]

[12] Wienke, A.; Ripatti, S.; Palmgrend, J.; Yashinf, A. (2009): *A bivariate survival model with compound Poisson frailty*, John Wiley & Sons

Nyilatkozat

Név: Török Anikó

ELTE Természettudományi Kar, szak: Biztosítási és pénzügyi matematika MSc

NEPTUN azonosító: XW72MI

Szakdolgozat címe: Túlélési modellek a biztosításban

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

Budapest, 2018. május 10.

_____

a hallgató aláírása