Biostatisztikai módszerek a rákkutatásban

Diplomamunka

Írta: Berki László

Alkalmazott matematikus szak

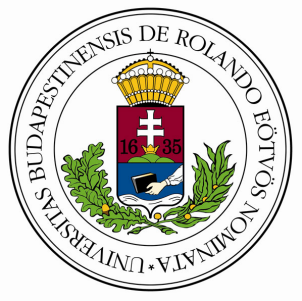
Témavezetők:

Móri Tamás, egyetemi docens

Valószínűségelméleti és Statisztika Tanszék

Tusnády Gábor, kutatóprofesszor

MTA Rényi Alfréd Matematikai Kutatóintézete, Országos Onkológiai Intézet



Eötvös Loránd Tudományegyetem

Természettudományi Kar

2010

Köszönetnyilvánítás

Ezúton szeretném megköszönni konzulensemnek, Móri Tamásnak, hogy a diplomamunka elkészülése alatt mindvégig segítette és felügyelte munkámat, sürgős esetekben bármikor szakított rám idejéből, és könyveit is korlátlan ideig használhattam. A tanulmány precízsége érdekében a benne szereplő apróbb hibákra is felhívta a figyelmemet, mind matematikaira, mind nyelvtanira, amiért külön hálás vagyok.

Továbbá szeretnék köszönetet mondani Tusnády Gábornak, aki nemcsak szakmai tanácsokkal látott el idejét nem sajnálva, de befogadott a kutatásával foglalkozó csoportba is, ezáltal megismerhettem munkásságát, munkatársait, és élőben lehettem tanúja a megbeszéléseknek.

Előszó

A túlélésanalízis, ami a diplomamunka témaköre, egy viszonylag új területe a statisztikának. A huszadik század második felében indult rohamos fejlődésnek, egyik legnagyobb kutatója D. R. Cox (1924-). A módszer neve és a vele kapcsolatos fogalmak arra utalnak, hogy elsősorban súlyos betegségek különböző kezeléseinek összehasonlítására alkalmazzák, ahol a vizsgált esemény a beteg halála, ill. annak időpontja a kezeléstől számítva. Általánosabban arra kereshetjük a választ, hogy egyes egyedek esetében miért nagyobb a kockázata a vizsgálat céljából fontos esemény bekövetkezésének.

Legfőbb célunk tehát, hogy egy beteg adataiból (ezeket a páciens paramétereinek fogom nevezni) túlélési valószínűséget tudjunk meghatározni. A könnyebb kezelhetőség szempontjából eloszlásfüggvénynek célszerű közismert függvényt vennünk, ezek közül a leggyakoribbakat a 2. fejezetben ismertetem. Az egyes paraméterek értékei közötti különbségek elemzéséhez modelleket alkalmazunk attól függően, hogy az adott paraméter milyen szintű változást visz végbe a beteg túlélési esélyeiben. Ezekről részletesen a 3. fejezetben lesz szó, ahol az ismert modellek mellett Tusnády Gábor saját fejlesztésű modellje is helyet kap. A különbségek számszerűsítése céljából regressziós módszereket alkalmazunk, amelyeket bővebben a 4. fejezetben tárgyalok. Végül az 5. fejezetben az SPSS és az R statisztikai programok segítségével valós adatokra alkalmazom az eddig leírtakat.

Az Egészségügyi Minisztérium 1998-ban létrehozta a Magyar Nemzeti Rákregisztert, ahol az ország összes regisztrált rákbetegének adatait nyilvántartják és felhasználják. Ennek oka, hogy Magyarországon három emberből kettőnél diagnosztizálnak élete során egyszer vagy többször valamilyen típusú rákot, közülük több százezren meg is halnak miatta, ezért alapvető feladatunk, hogy matematikailag modellezzük a különböző rákbetegségeket.

Tartalomjegyzék

1. Bevezetés

Az orvosi statisztikában a leggyakoribb vizsgálatok tárgya egy esemény (általában a halál, de lehet a tünet enyhülése, felépülés stb.) bekövetkezésének ideje, ezt ezentúl meghibásodási időnek fogjuk hívni. Az ilyen típusú adatokat nevezzük élettartam-adatoknak, ezek elemzésével foglalkozik a túlélésanalízis. A statisztika ezen területét nem csupán az alkalmazásbeli sajátossága miatt különítik el, hanem mert a hagyományos statisztikai módszerekkel nehezen kezelhetők az élettartam-adatok. Ennek egyik oka, hogy ezek általában *pozitívan ferdék*, vagyis viszonylag kevés a nagy érték, vagyis például a normális eloszlás feltételezése itt nem helytálló.

A másik, és egyben a nagyobb probléma, hogy a túlélési idők gyakran *cenzoráltak*. Egy páciens túlélési idejét cenzoráltnak mondjuk, ha a meghibásodási időt nem ismerjük. Ez több okból megtörténhet, leggyakoribb, hogy a páciens a megfigyelés végén még életben volt, vagy hogy a megfigyelés alatt meghalt, de nem a vizsgált betegség miatt. A cenzorálás vizsgálata nem csak azért fontos, mert ezt figyelmen kívül hagyva csökken az információnk, hanem mert torzul is az eloszlásunk, hiszen például ha egy 10 éves vizsgálat alatt az 1000 páciensből 700 túléli a betegséget, és őket nem vesszük figyelembe, akkor azt kapjuk, hogy senki sem éli túl a 10 évet, miközben tudjuk, hogy 700 beteg nem halt bele. *Jobb oldali cenzorálásról* akkor beszélünk, ha a megfigyelési időnél többet élt (volna), ha kevesebbet, akkor bal oldaliról. Utóbbi csak ritkán fordul elő, ezért mostantól cenzorálás alatt jobb oldali cenzorálást fogok éretni. Ezáltal az eredeti D_i $(i = 1, \dots, n)$ minta helyett az (T_i, δ_i) $(i = 1, \dots, n)$ ún. *cenzorált mintával* fogunk dolgozni, ahol $T_i = D_i \wedge C_i$, C_i az i-edik cenzorálási idő és $\delta_i = \chi_{\{T_i = D_i\}}$, tehát azt mutatja, hogy az i-edik páciens meghalt vagy cenzoráltuk. Két speciális sémát emelünk ki, az 1. és 2. típusú cenzorálást. Az 1. típusúnál minden cenzorálási idő determinisztikus és azonos ($C_i = c \in \mathbb{R}^+$, $i = 1, \dots, n$), míg a 2. típusúnál az első s meghibásodást várjuk meg, azaz $C_i = D_s^*$ $(i = 1, \dots, n)$. Előbbinél c tipikus esete a vizsgálat vége, így végig a cenzorálás ezen fajtájával fogunk foglalkozni.

A megszokott statisztikai módszereknél a mintánk többnyire független, azonos eloszlású. A függetlenség itt is teljesül, de az eloszlások egyedről-egyedre változhatnak, például közismert, hogy a mellrák jóval veszélyesebb a nőknél, mint a férfiak esetében, vagyis két különböző nemű páciens élettartam eloszlása biztosan különbözik. Azokat a változókat, amelyek befolyásolják az eloszlást, mint például a kor, nem, dohányzás, *magyarázó változóknak* nevezzük, egy konkrét beteg esetén *paramétereknek*. Ha két páciens minden paramétere megegyezik, akkor a két páciens élettartam eloszlását azonosnak tekintjük.

A tanulmány során az élettartamokra elsősorban nem az eloszlásfüggvényükkel fogunk hivatkozni, hanem annak valamilyen függvényével. Ilyen például a *túlélésfügg-*

vény, ami a T valószínűségi változó esetén:

$$\overline{F}(t) := P(T \geq t) = 1 - F(t),$$

vagyis annak a valószínűsége, hogy az egyed megéli a t időt. Megjegyzem, hogy mivel a valószínűségi változó értéke idő, ezért természetesen csak olyan eloszlásokkal foglalkozunk, amelyek a nemnegatív számhalmazra koncentráltak. Másik fontos függvény a *kumulált hazárdfüggvény*, ami definíció szerint:

$$R(t) := -\log \overline{F}(t).$$

A felírásból látszik, hogy $R(0) = 0$, R monoton nő és $\lim_{t\to+\infty} R(t) = +\infty$. Ha az eloszlás abszolút folytonos, akkor T-nek létezik sűrűségfüggvénye, jelölje ezt f. Ekkor T *hazárdfüggvényén* az

$$r(t) := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} P(T < t + \varepsilon \mid T \geq t)$$

kifejezést értjük. Ez tulajdonképpen azt mutatja, hogy ha valaki a t időpontban él, akkor mekkora "valószínűséggel" fog ott meghalni, ezért néha szokás a túlélés *intenzitásának* nevezni (a szó szoros értelemben véve persze nem valószínűség, hiszen a hazárdfüggvény értékkészlete $\mathbb{R}^+$). Egyszerű számolással könnyen igazolható, hogy:

$$r(t) = R'(t) = \frac{f(t)}{\overline{F}(t)},$$

amiből R tulajdonságai miatt következik, hogy $r(t) \geq 0\ \forall t \geq 0$-ra.

Mint azt az előszóban már említettem, elsősorban a páciensek túléléseire szeretnénk becslést adni. Ezt túlnyomórészt az *általánosított maximum likelihood módszerrel* (Kiefer és Wolfowitz, 1956) tesszük meg, miszerint az $(\Omega, \mathcal{A})$ mérhető téren értelmezett $\mathcal{P} = \{P\}$ eloszláscsaládban $\hat{P}$ *általánosított maximum likelihood becslés*, ha $\forall P \in \mathcal{P}$:

$$\frac{d\hat{P}}{d\left(\hat{P} + P\right)}(X) \geq \frac{dP}{d\left(\hat{P} + P\right)}(X),$$

ahol X az adott minta. Jelölje $0 < a_1 < a_2 < ... < a_d$ a különböző meghibásodási időket, m_i az a_i multiplicitását, $n_i = |\mathcal{R}(a_i)|$, ahol $\mathcal{R}(t)$ a "*risk set*", a t-ben megfigyelés alatt álló egyedek halmaza, azaz $\mathcal{R}(t) = \{i : X_i \geq t\}$. Az általánosított ML-ból kiindulva determinisztikusan cenzorált mintából a Kaplan-Meier becsléssel

(1958) közelíthetjük a túlélésfüggvényünket, mégpedig a következőképpen:

$$\widehat{\overline{F}}(t) = \prod_{i:\, a_i<t} \left(1 - \frac{m_i}{n_i}\right).$$

Ennek a szórásnégyzetét a Greenwood-formula (1926) segítségével becsülhetjük:

$$D^2\left(\widehat{\overline{F}}(t)\right) \approx \left(\widehat{\overline{F}}(t)\right)^2 \sum_{i:\, a_i<t} \frac{m_j}{n_j\,(n_j - m_j)}.$$

Általánosabban, nézzük meg, hogyan írható fel a likelihood-függvény cenzorált minta esetén. Cenzorálatlan esetben a hagyományos likelihood-függvény:

$$L = \prod_{i=1}^{n} f(a_i).$$

Most tegyük fel, hogy az n megfigyelésből d egyed meghal az $a_1, \ldots, a_d$ időpontokban, a maradék $n-d$ egyedet pedig a $c_1, \ldots, c_{n-d}$ időkben jobbról cenzoráljuk. Ha valakit c-ben cenzorálunk, az azt jelenti, hogy legalább c időt él, aminek a valószínűsége $\overline{F}(c)$. Ezt felhasználva, cenzorált mintára a likelihood-függvény:

$$L = \prod_{i=1}^{d} f(a_i) \prod_{j=1}^{n-d} \overline{F}(c_j).$$

Olvasszuk egybe a meghibásodási és cenzorálási időket, rendeljük az i-edik egyedhez az (t_i, δ_i) párt, ahol δ_i a meghibásodás indikátora. Ezzel:

$$L = \prod_{i=1}^{n} \{f(t_i)\}^{\delta_i} \left\{\overline{F}(t_i)\right\}^{1-\delta_i} = \prod_{i=1}^{n} \{r(t_i)\}^{\delta_i}\, \overline{F}(t_i). \tag{1}$$

Az elemi statisztikából ismert eloszláscsaládokon kívül érdemes definiálnunk néhány, speciális tulajdonsággal rendelkező eloszláscsaládot, amik segítségével jobb becslést tudunk adni az eloszlásra (feltéve, hogy az vizsgált élettartam eloszlás az adott eloszláscsalád tagja). Ilyen eloszláscsalád IFR, DFR, IFRA, DFRA, NBU és NWU.

1. Definíció. [Increasing/Decreasing Failure Rate] *A T valószínűségi változóhoz tartozó F eloszlásfüggvényre $F \in$IFR ($F \in$DFR), ha $\forall s > 0$-ra az $\frac{\overline{F}(t+s)}{\overline{F}(t)}$ t-nek monoton fogyó (növekvő) függvénye.*

1. Tétel. *$F \in$IFR ($F \in$DFR) $\Longleftrightarrow$ R konvex (konkáv). Ha $\exists f$, akkor ekvivalens r monoton növő (csökkenő) voltával.*

2. Definíció. [Increasing/Decreasing Failure Rate Average] *A T valószínűségi változóhoz tartozó F eloszlásfüggvényre $F \in$IFRA ($F \in$DFRA), ha az $\overline{F}(t)^{1/t}$ t-nek*

monoton fogyó (növekvő) függvénye.

2. Tétel. *$F \in$IFRA ($F \in$DFRA) $\Longleftrightarrow$ $R(t)/t$ monoton növő (csökkenő) t-ben.*

3. Definíció. [New Better/Worse than Used] *A T valószínűségi változóhoz tartozó F eloszlásfüggvényre $F \in$NBU ($F \in$NWU), ha $\forall t, s > 0$-ra $\overline{F}(t+s) \leq (\geq)\overline{F}(t)\overline{F}(s)$.*

3. Tétel. *$F \in$NBU ($F \in$NWU) $\Longleftrightarrow$ R szuperadditív (szubadditív).*

A definíciókban szereplő függvényeket valószínűségekké átírva kapjuk, hogy az IFR családban minél idősebb az egyed, annál rosszabbak az életkilátásai, míg az NBU szemléletesen azt jelenti, hogy az új egyed életkilátásai jobbak, mint azé, aki már élt valamennyit. Ezen osztályok között fennáll a IFR$\subsetneqq$IFRA$\subsetneqq$NBU és a DFR$\subsetneqq$DFRA$\subsetneqq$NWU tartalmazás, ezek bizonyítása triviális.

Gyakori feladat a rákkutatás tanulmányozása során, hogy mielőtt páciensek két csoportját összehasonlítanánk az élettartamuk szempontjából, eldöntsük, egyáltalán különböznek-e. Ennek tesztelésére az egyik lehetséges módszer a *log-rang teszt* (Mantel-Haenszel, 1958), amely a következőképpen jár el: tegyük fel, hogy a két csoportban összesen d különböző meghibásodási idő van, ezek rendre $0 < a_1 < \ldots < a_d$. Legyen a j-edik ($j = 1, 2$) csoportban az a_i-ben meghalt egyedek száma m_{ij}, a megfigyelés alatt állóké n_{ij}, továbbá $m_i := m_{i1} + m_{i2}$, $n_i := n_{i1} + n_{i2}$. Ekkor:

$$\frac{U^2}{V} \sim \chi_1^2,$$

ahol:

$$U = \sum_{i=1}^{d} \left(m_{i1} - \frac{n_{i1} m_i}{n_i} \right),$$

$$V = \mathrm{var}(U) = \sum_{i=1}^{d} \frac{n_{i1} n_{i2} m_i (n_i - m_i)}{n_i^2 (n_i - 1)}.$$

2. Eloszlások

A túlélések vizsgálatakor számos eloszlással kísérleteznek, természetesen (az időskála miatt) csak azokkal, melyek a nemnegatív valós számokon vannak értelmezve. Ide tartozik az exponenciális, a Weibull, a Gompertz-Makeham, a lognormális, a log-logisztikus, a gamma és az inverz Gauss eloszlás is. Ezek közül részletesen csak az exponenciálissal, a Weibullal és a lognormálissal foglalkozom.

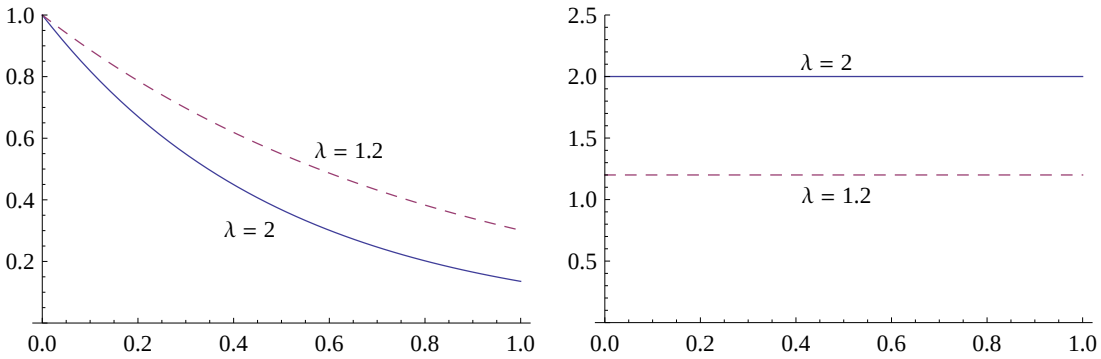
2.1. Exponenciális eloszlás

Túlélésanalízisben a legkönnyebben kezelhető eloszlás az *exponenciális eloszlás*, melyet egyszerűségéért használnak szívesen. A $\lambda > 0$ paraméterű exponenciális eloszlás túlélésfüggvénye:

$$\overline{F}(t) = e^{-\lambda t},$$

ill. hazárdfüggvénye:

$$r(t) = \lambda.$$



(a) Túlélésfüggvény (b) Hazárdfüggvény

1. ábra. *Exponenciális eloszlás*

Mint látjuk, ez egy egyparaméteres eloszlás, konstans hazárdfüggvénnyel, vagyis a túlélés intenzitása nem függ az időtől, mindvégig állandó. Az eloszlás várható értéke $1/\lambda$, szórásnégyzete $1/\lambda^2$. A paraméter ML-becslése a legtermészetesebb módon történik: a meghibásodások átlagának reciproka. Mindezen jó tulajdonságok ellenére sajnos ritkán fordul elő az orvostudományban.

2.2. Weibull eloszlás

Az exponenciális eloszlással ellentétben a *Weibull eloszlás* már jóval gyakoribb, köszönhető ez többek között két paraméterének. Az X valószínűségi változó Weibull eloszlású

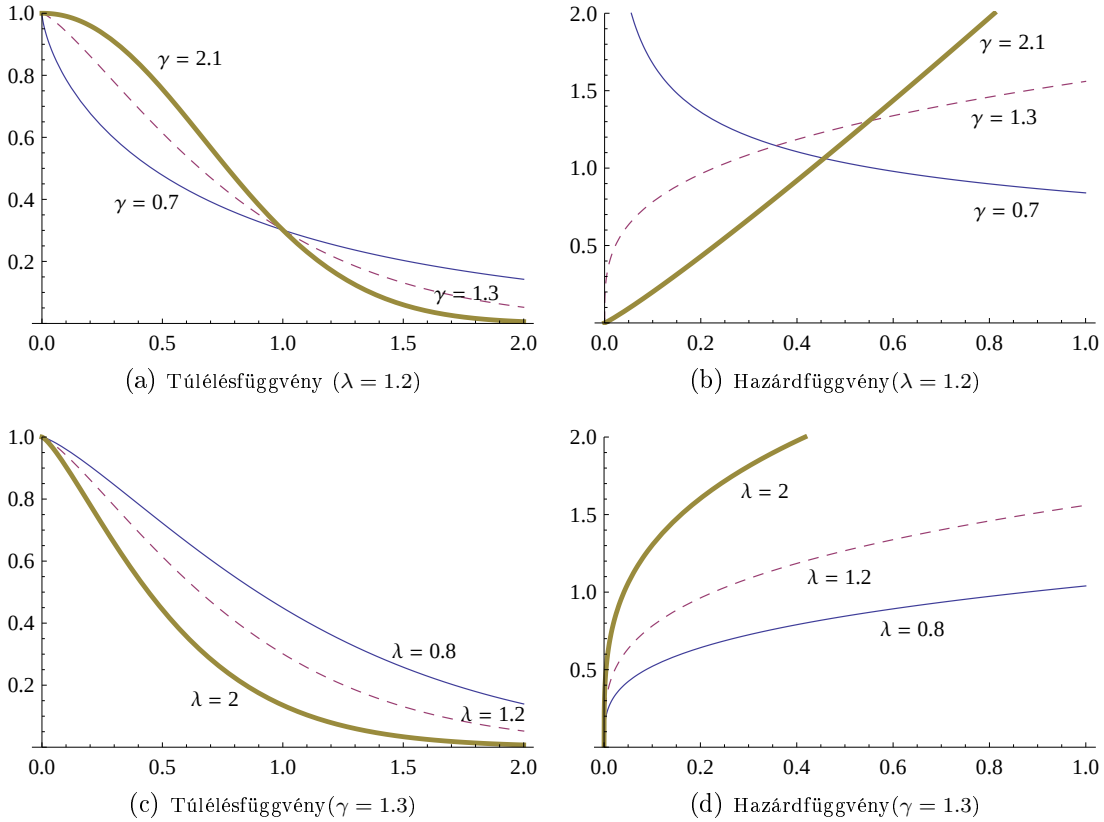
$\lambda > 0$ és $\gamma > 0$ paraméterekkel ($X \sim W(\lambda, \gamma)$), ha túlélésfüggvénye:

$$\overline{F}(t) = e^{-\lambda t^{\gamma}},$$

hazárdfüggvénye:

$$r(t) = \lambda \gamma t^{\gamma - 1}.$$

ahol λ-t skálaparaméternek, γ-t pedig alakparaméternek nevezik, amely elnevezéseket a 2. ábra jól szemlélteti. Speciálisan, ha $\gamma = 1$, akkor az exponenciális eloszlást kapjuk. γ egyéb értékeire a hazárdfüggvény szigorúan monoton növő (tehát $F \in IFR$), ill. csökkenő ($F \in DFR$) attól függően, hogy γ 1-nél nagyobb vagy kisebb.



(a) Túlélésfüggvény ($\lambda = 1.2$) (b) Hazárdfüggvény($\lambda = 1.2$)

(c) Túlélésfüggvény($\gamma = 1.3$) (d) Hazárdfüggvény($\gamma = 1.3$)

2. ábra. *Weibull eloszlás*

A Weibull eloszlás várható értéke $EX = \lambda^{-1/\gamma}\Gamma\left(\frac{\gamma+1}{\gamma}\right)$, szórásnégyzete pedig $D^2(X) = \lambda^{-2/\gamma}\left[\Gamma\left(\frac{\gamma+2}{\gamma}\right) - \Gamma\left(\frac{\gamma+1}{\gamma}\right)^2\right]$. A paraméterek ML-becslését a következő egyenletrendszer megoldása adja:

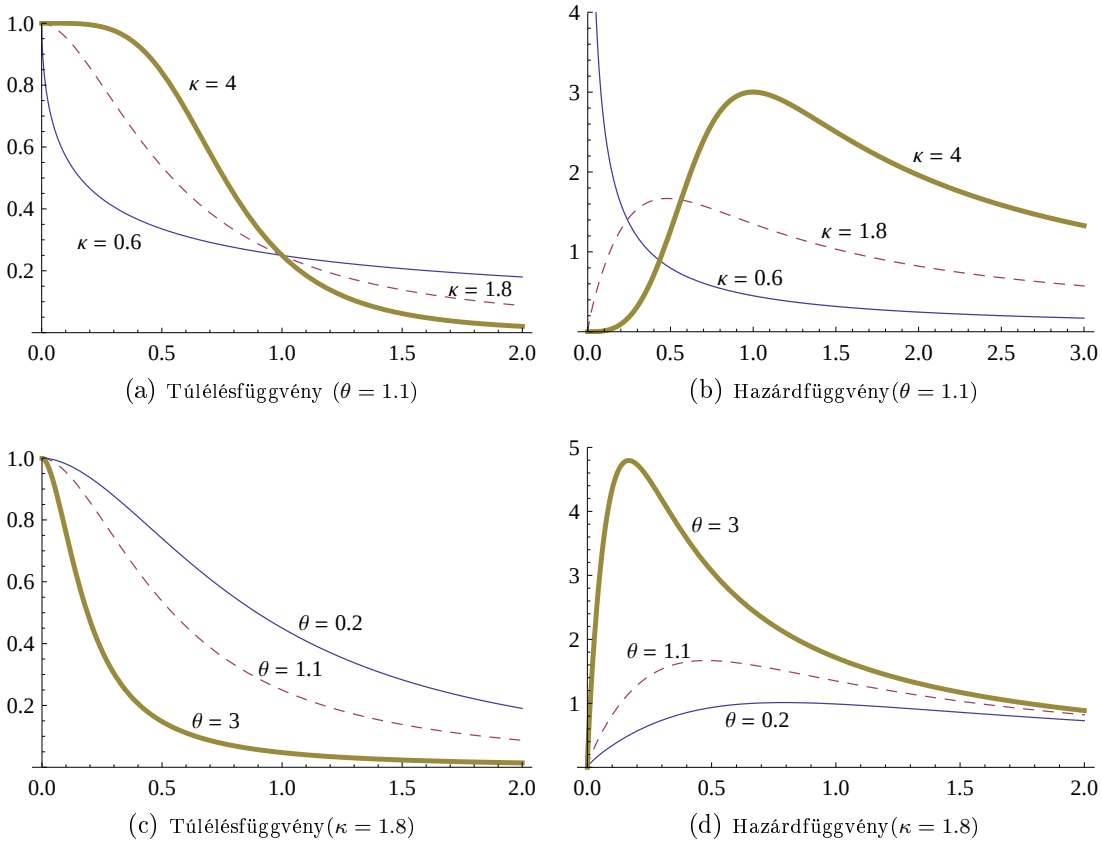
$$\frac{d}{\widehat{\gamma}} + \sum_{i=1}^{n} \delta_i \log a_i - \frac{d}{\sum_i a_i^{\widehat{\gamma}}} \sum_{i=1}^{n} a_i^{\widehat{\gamma}} \log a_i = 0$$

$$\widehat{\lambda} = d/\sum_{i=1}^{n} a_i^{\widehat{\gamma}}.$$

Ez explicit módon nem számolható ki, csak numerikus módszerekkel tudjuk közelíteni, például a Newton-Raphson-módszerrel (ld. Függelék).

2.3. Log-logisztikus eloszlás

A Weibull eloszlás egyik hátránya, hogy tetszőleges paraméterezés mellett a hazárdfüggvény monoton, bár egyes kezeléseknél előfordul, hogy az intenzitás egy ideig nő (amíg a szervezet befogadja az új hatóanyagot), aztán csökken (a szer elkezd gyógyítólag hatni a páciensre).



(a) Túlélésfüggvény ($\theta = 1.1$) (b) Hazárdfüggvény($\theta = 1.1$)

(c) Túlélésfüggvény($\kappa = 1.8$) (d) Hazárdfüggvény($\kappa = 1.8$)

3. ábra. *Log-logisztikus eloszlás*

A *log-logisztikus eloszlás* túlélésfüggvénye, ill. hazárdfüggvénye:

$$\overline{F}(t) = \frac{1}{1 + e^{\theta} t^{\kappa}},$$

$$r(t) = \frac{e^{\theta} \kappa t^{\kappa - 1}}{1 + e^{\theta} t^{\kappa}}.$$

Ezt az eloszlást jelöljük $LLog(\theta, \kappa)$-val. Az elnevezés onnan származik, hogy ha X log-logisztikus eloszlású, akkor $\log X$ logisztikus eloszlású.

A hazárdfüggvény $\forall \kappa \geq 1$-re monoton csökkenő, $0 < \kappa < 1$-re egymódusú, pontosabban létezik (globális) maximuma (ld. 3. ábra). A várható érték $e^{-\theta/\kappa}\kappa/\sin\kappa$, a szórásnégyzet $e^{-2\theta/\kappa}\left(\frac{2\kappa}{\sin 2\kappa} - \frac{\kappa^2}{\sin^2\kappa}\right)$.

3. Modellek

A túlélések vizsgálatakor az egyik legfontosabb feladatunk megállapítani, hogy milyen tényezők milyen irányban befolyásolják a páciensek túléléseit (természetszerűleg az életkor növekedtével romlanak a túlélési valószínűségek). Szeretnénk elérni, hogy egy páciens paramétereiből túlélésfüggvényt tudjunk generálni. Ehhez persze ismernünk kell a faktorok egy konkrét értéke melletti túlélésfüggvényt, ezt nevezzük alap túlélésfüggvénynek, jelöljük ezt $\overline{F}_0$-lal. Ebből kiindulva a tényezők módosításával kapjuk egy más paraméterezésű páciens túlélésfüggvényét.

3.1. Gyorsított modellek

A vizsgálatok során gyakran előfordul, hogy egy kezelés hatását szeretnénk igazolni, vagy egy új kezelést összevetni az eddig használttal. Ilyen esetekben célszerű gyorsított modelleket alkalmazni, összehasonlítva a két csoportot (kezelt-nem kezelt/új módszer-régi módszer). Fontos megjegyeznünk, hogy az egyes páciensek véletlenszerűen kerülnek a kezelt, ill. a kontrollcsoportba. Legyen a kontrollcsoport túlélésfüggvénye $\overline{F}_0(t)$, a kezelté $\overline{F}_1(t)$. Ekkor a *gyorsított élet modell* szerint:

$$\overline{F}_1(t) = \overline{F}_0(bt),$$

ahol b a gyorsító paraméter. Vezessük be a $b = e^\beta$ jelölést, ezzel az előző képlet:

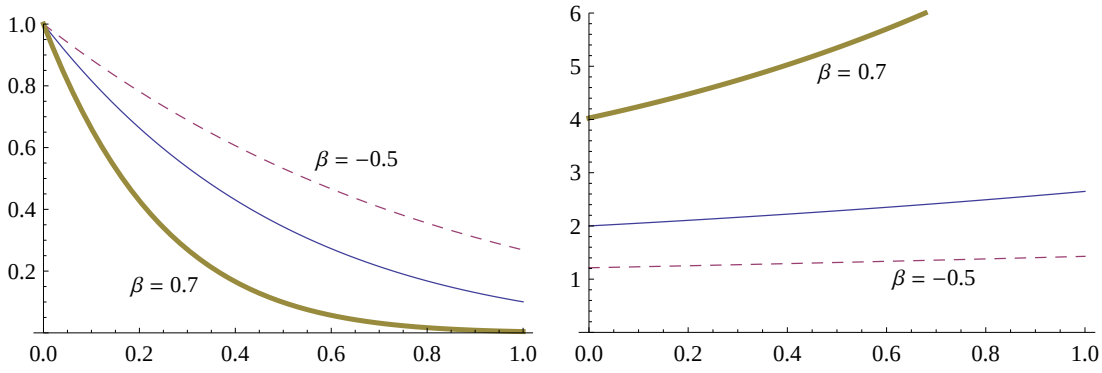
$$\overline{F}_1(t) = \overline{F}_0(e^\beta t),$$

vagy a hazárdfüggvényekre vonatkozó alakban:

$$r_1(t) = r_0(e^\beta t)e^\beta. \tag{2}$$

A gyorsított élet modell tehát nem más, mint a túlélésfüggvény időskála szerinti multiplikatív módosítása, vagyis a betegség sebességének változtatása. Ha $\beta = 0$, akkor a két csoport nem különbözik egymástól, $\beta > 0$ esetén károsabb az új módszer a réginél, $\beta < 0$ mellett pedig előnyös, például ha $\beta = -\log 2$, akkor a kezelés hatására az életkilátásaink romlásának sebességét felezi, vagyis hasznos.

Egy (új) kezelés a kezdeti ($t = 0$) időpontban még nem fejti ki hatását a betegre, így egy logikus követelmény a modellel szemben, hogy a hazárdfüggvények ebben az időpontban megegyezzenek, vagyis hogy $r_1(0) = r_0(0)$ teljesüljön. (2) miatt viszont $r_1(0) = r_0(0)e^\beta$, amiből kapjuk, hogy $\beta = 0$, vagyis a két túlélésfüggvény minden pontban megegyezik, így ez a feltétel erre a modellre nem teljesül.



(a) Túlélésfüggvény (b) Hazárdfüggvény

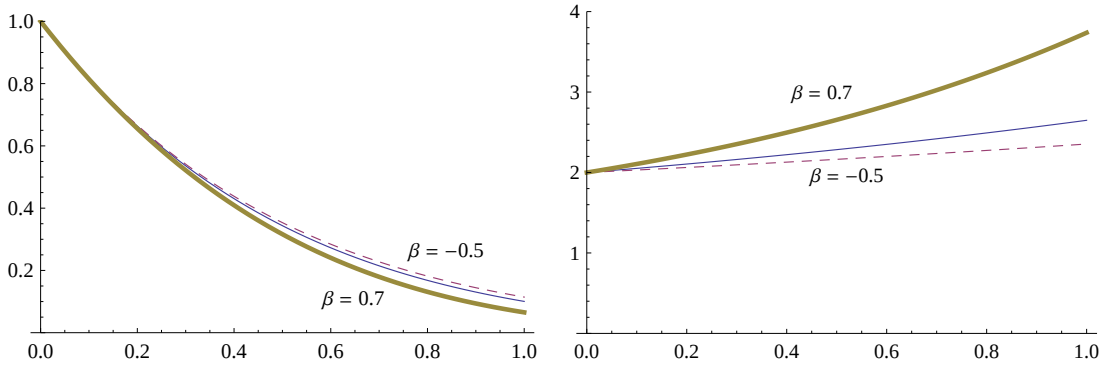
4. ábra. *Gyorsított élet modell*

Ennek elkerülése érdekében vezessünk be egy új modellt, a *gyorsított hazárd modellt* (Chen és Wang, 2000). Ennél a modellnél a túlélésfüggvény helyett a hazárdfüggvényt gyorsítjuk, ez alapján a kezelt csoport hazárd-, ill. túlélésfüggvénye:

$$r_1(t) = r_0(e^{\beta}t),$$

$$\overline{F}_1(t) = \left[\,\overline{F}_0(e^{\beta}t)\,\right]^{e^{-\beta}}.$$

A gyorsított élet modellhez hasonlóan itt is β előjele dönti el a kezelés hatását, itt viszont $\beta = -\log 2$ esetén a túlélés intenzitásváltozásának sebessége csökken a felére. Ennél a modellnél már tudjuk teljesíteni az $r_1(0) = r_0(0)$ feltételt anélkül, hogy a két csoport hazárd- és túlélésfüggvénye megegyezzen, sőt, tetszőleges $\beta \neq 0$ mellett teljesül a feltétel (ld. 5. ábra).



(a) Túlélésfüggvény (b) Hazárdfüggvény

5. ábra. *Gyorsított hazárd modell*

Ezen túl a gyorsított hazárd modell rendelkezik még egy hasznos tulajdonsággal, amit a következő tételben bizonyítok.

4. Tétel. *Ha $\beta < 0$ és a hazárdfüggvény szigorúan monoton növő/csökkenő, akkor a kezelés előnyös/hátrányos. Ha $\beta > 0$ és a hazárdfüggvény szigorúan monoton növő/csökkenő, akkor a kezelés hátrányos/előnyös.*

Bizonyítás. Csak a $\beta < 0$, szigorúan monoton növő esetre látom be, a többi hasonlóképpen igazolható.

Ha $\beta < 0$, akkor $e^{\beta} < 1$, így a hazárdfüggvény monoton növekedése miatt:

$$r_0(s) > r_0(e^{\beta}s) = r_1(s).$$

Ez $\forall s \geq 0$-ra teljesül, ezért $\forall t \geq 0$ esetén:

$$R_0(t) = \int_0^t r_0(s)ds > \int_0^t r_1(s)ds = R_1(t).$$

Ebből:

$$\overline{F}_0(t) = e^{-R_0(t)} < e^{-R_1(t)} = \overline{F}_0(t),$$

tehát a kezelés tényleg hasznos. □

Adott minta esetén szeretnénk becsülni $b = e^{\beta}$-t (r_0-t ismertnek feltételezve), erre Chen és Wang [9] adott egy numerikus algoritmust, melynek lépései a következők:

1. Legyen b_0 egy tetszőleges pozitív, valós szám. Az eredeti meghibásodási időket $(a_i,\ i = 1, \ldots, n)$ szorozzuk meg $b_0^{x_i}$-nel, ahol x_i a kezelés indikátora.

2. Alkalmazzuk a parciális likelihood becslés módszerét a módosított adatokra az alkalmas arányos hazárd modell (ld. 4.2. és 3.2. szakasz) megtalálására, azaz:

$$r(t \mid X = x_i, b) = r_0(t)b^{-x_i} \quad (i = 1, \ldots, n),$$

   ahol X a magyarázó változó. Jelölje itt b becslését $\hat{\psi}(b_0)$.

3. Ismételjük az 1. és 2. lépést addig, amíg nem találunk egy olyan $\tilde{b}$-ot, melyre vagy $\hat{\psi}(\tilde{b}) = \tilde{b}$, vagy $\left[\hat{\psi}(\tilde{b}+0) - \tilde{b}\right] \cdot \left[\hat{\psi}(\tilde{b}-0) - \tilde{b}\right] \leq 0$.

Eddig az időfüggetlen esettel foglalkoztunk, most térjünk át az időtől függőre, amelyet Finkelstein [10] vizsgált részletesen. Vegyük alapul a gyorsított élet modellt, amely időfüggő esetben:

$$\overline{F}_1(t) = \overline{F}_0\left(\beta(t)\right), \tag{3}$$

ahol $\beta(t)$ az időfüggő skála-transzformációs függvény, ami a páciens relatív öregedését méri ($\beta(t) = bt$ esetén az időfüggetlen változatot kapjuk, ha $\beta(t) > t$, $\forall t \geq 0$, akkor

a kezelés káros hatású, ugyanis $\overline{F}_1(t) = \overline{F}_0(\beta(t)) < \overline{F}_0(t)$; analóg módon, $\beta(t) < t$ teljesülése esetén a kezelés jótékony). $\overline{F}_1$ pontosan akkor lesz eloszlásfüggvény, ha $\beta(0) = 0$, β monoton növő és $\lim_{t\to+\infty} \beta(t) = +\infty$. Tegyük fel, hogy β-ra teljesül ez a három feltétel, valamint hogy differenciálható a $[0, +\infty)$ intervallumon, így:

$$\beta(t) = \int_0^t \varphi(s)ds.$$

Ekkor β monotonitása miatt $\varphi \geq 0$. (3)-at a kumulált hazárdfüggvényekkel felírva:

$$R_1(t) = R_0\left(\beta(t)\right). \tag{4}$$

Az $R(t)$ kumulált hazárdfüggvény szigorúan monoton, folytonos függvény, így $\exists R^{-1}(t)$, ezt felhasználva:

$$\beta(t) = R_0^{-1}\left(R_1(t)\right).$$

(4)-et deriválva kapjuk a hazárdfüggvényekre vonatkozó összefüggést:

$$r_1(t) = r_0\left(\beta(t)\right)\varphi(t).$$

A gyorsított modellek tehát az időskálát paraméterezik át, a gyorsított élet a túlélésfüggvényét, a gyorsított hazárd a hazárdfüggvényét, vagyis ezekkel a modellekkel a túlélés sebességét változtathatjuk igazolva, vagy éppen megcáfolva egy kezelés hatásosságát.

3.2. Arányos hazárd modell

A gyorsított modellekhez hasonlóan, az *arányos hazárd modellel* is két csoport, azaz két különböző paraméterezésű páciens túléléseit kívánjuk összevetni. Ennél a modellnél (nevéből adódóan) a hazárdfüggvények arányát feltételezzük az idő függvényében konstansnak, tehát:

$$\frac{r_1(t)}{r_0(t)} \equiv b > 0,$$

ahol r_1 és r_0 a két páciens hazárdfüggvénye, $b \in \mathbb{R}^+$pedig ezen két beteg paramétereitől függő konstans. Ennek a feltételnek következménye, hogy a túlélésfüggvények nem keresztezhetik egymást. Általánosan, jelöljük $\underline{X} = (X_1, \ldots, X_k)^\top$-val a magyarázó változók által meghatározott vektorváltozót, és tegyük fel, hogy minden koordinátája indikátorváltozó. Ekkor az $\underline{X} = \underline{x}$ paraméterű páciens hazárd- és túlélésfüggvénye

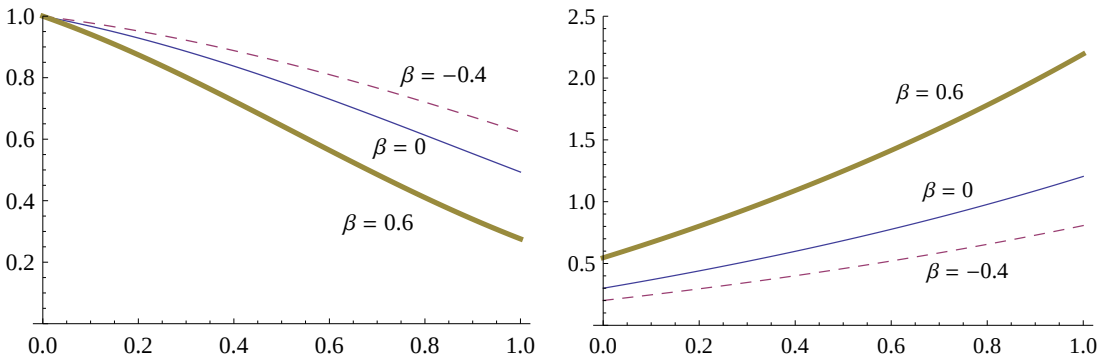
felírható speciális alakban:

$$r_{\underline{x}}(t) = r_0(t)\psi(\beta_1 x_1 + \ldots + \beta_k x_k),$$

$$\overline{F}_{\underline{x}}(t) = \left[\overline{F}_0(t)\right]^{\psi(\beta_1 x_1 + \ldots + \beta_k x_k)}.$$

ahol $r_0(t)$ az alap hazárdfüggvény, vagyis az $\underline{x} = (0, 0, \ldots, 0)^\top$ paraméterezés melletti páciens hazárdfüggvénye, $\beta_i \in \mathbb{R}$ $(i = 1, \ldots, k)$ az i-edik magyarázó változó együtthatója, ψ pedig egy ismert pozitív függvény (relatív hazárd), amire $\psi(0) = 1$. Ezen tulajdonságai miatt általában a $\psi(x) = e^x$ függvénnyel szoktak dolgozni, ezt a modellt hívjuk *Cox-modellnek* (vagy loglineáris modellnek), ahol a hazárdfüggvény tehát:

$$r_{\underline{x}}(t) = r_0(t)e^{\beta_1 x_1 + \ldots + \beta_k x_k}. \tag{5}$$



(a) Túlélésfüggvény (b) Hazárdfüggvény

6. ábra. *Arányos hazárd modell*

Itt az $\eta = \beta_1 x_1 + \ldots + \beta_k x_k$ kifejezést a páciens *prognózis-indexének* nevezzük, ennek ismeretében már fel tudjuk írni a túlélésfüggvényét. A gyorsított modellekhez hasonlóan, ha $\eta > 0$, akkor a kontrollcsoport túlélése a jobb, ha $\eta < 0$, akkor a kezelt csoporté.

A továbbiakban csak a Cox-modellel foglalkozunk. Ez visszavezethető a lineáris modellre, hiszen (5)-öt átalakítva kapjuk, hogy:

$$\log\left(\frac{r_{\underline{x}}(t)}{r_0(t)}\right) = \beta_1 x_1 + \ldots + \beta_k x_k.$$

A magyarázó változókról feltettük, hogy indikátorváltozók, ami persze nem mindig teljesül. Legyen M egy ν különböző értéket felvevő (ν szintű) valószínűségi változó. M előállítható $\nu - 1$ indikátorváltozóval, legyenek ezek $X_{M2}, X_{M3}, \ldots, X_{M\nu}$. A megfeleltetést az alábbi táblázat definiálja:

M szintjei	X_{M2}	X_{M3}	...	$X_{M\nu}$
1.	0	0	...	0
2.	1	0	...	0
3.	0	1	...	0
...	...	...	...	...
ν.	0	0	...	1

Ritkán abszolút folytonos magyarázó változó is előfordulhat, ezt diszkretizálva és a fenti módszert alkalmazva szintén azonosítani tudjuk véges számú indikátorváltozóval. Adódhat azonban olyan helyzet is, amikor két vagy több változó együttes jelenléte (vagy hiánya) felerősíti vagy legyengíti a hatást, például a gégeráknál tapasztalták, hogy az ösztrogénhiányosok túlélési esélyei jóval rosszabbak, mint a tesztoszteronhiányos férfiaknak. Így ha X_1 jelöli a nemet (0: férfi, 1: nő), X_2 a nemi hormon mennyiségét (0: normál vagy több, 1: kevés), akkor a két magyarázó változó kölcsönhatását a következőképpen tudjuk vizsgálni:

$$r_{\underline{x}}(t) = r_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2}.$$

Most már tetszőleges időfüggetlen magyarázó változót tudunk kezelni. Orvosilag és technikailag is fontos kérdés, hogy ezek közül melyek a szignifikánsak. Tegyük fel, hogy adott két különböző modellünk (A-modell és B-modell) ugyanarra a mintára úgy, hogy az A-modell magyarázó változói (p db) valódi részhalmazát képezik a B-modell magyarázó változói (q db) halmazának, vagyis az A-modell paraméteresen bele van ágyazva a B-modellbe. Jelölje $\hat{L}_A$, ill. $\hat{L}_B$ a két modell maximalizált likelihood-függvényét. Alkalmazzuk az illeszkedésvizsgálatnál használt likelihood-hányados statisztikát mindkét modellre, és vegyük ezek különbségét:

$$-2\log\left(\frac{\hat{L}_A}{\hat{L}_B}\right).$$

Ennek segítségével tesztelhetjük a modellt $q-p$ szabadságfokú χ^2-próbával azon nullhipotézis mellett, hogy a csak a B-modellben szereplő magyarázó változókhoz tartozó együtthatók értéke 0, tehát hogy ezek a magyarázó változók nem szignifikánsak. Ezt

Időfüggőség mellett annyival módosul a modellünk, hogy a magyarázó változók nem konstansok, hanem ismert függvények. Ezen belül megkülönböztetünk *belső és külső változókat*, ahol a belső változót csak életben lévő páciensnél tudunk mérni (pl. vérnyomás), a külső változóhoz pedig nem szükséges, hogy a beteg életben legyen (pl.

kor). Ezzel a Cox-modell időfüggő alakja:

$$r_{\underline{x}(t)}(t) = r_0(t)e^{\beta_1 x_1(t)+\ldots+\beta_k x_k(t)}.$$

3.3. A modellek tulajdonságai

A megfelelő modell kiválasztásához számos dolgot számításba kell vennünk, többek között az alapeloszlást, a különböző paraméterű páciensekhez tartozó eloszlás- és hazárdfüggvények viszonyát, de a különbözőség ismert biológiai hatása is segítségünkre lehet. Például ha a kezelés hatékony, és a hatás nagyjából állandó, akkor az arányos hazárd vagy a gyorsított élet kézenfekvőbb, ha viszont fokozatosan erősödik, akkor a gyorsított hazárd modellt célszerű alkalmazni.

Ha ismerjük az alapeloszlást, ami egy adott eloszláscsalád tagja, és szeretnénk ezen eloszláscsaládon belül maradni, akkor ez a kikötésünk korlátozza a szóba jövő modelleket. Egyetlen kivétel van, aminél bármelyik modell alkalmazása után ugyanabban az eloszláscsaládban maradunk, ezt a következő tételben igazolom, ami bizonyítás nélkül [9]-ben megtalálható.

5. Tétel. *Időfüggetlen magyarázó változók esetén a gyorsított élet, gyorsított hazárd, ill. arányos hazárd modellek pontosan akkor ekvivalensek, ha az alapeloszlás Weibull.*

Bizonyítás. Az elégségesség egyszerűen igazolható, alkalmazzuk mindhárom modellt a $W(\lambda, \gamma)$ alapeloszlásra, azaz $r_0(t) = \lambda\gamma t^{\gamma-1}$.
1. Gyorsított élet: $r_X(t) = r_0(te^\beta)e^\beta = \lambda\gamma\left(te^\beta\right)^{\gamma-1}e^\beta = \lambda\gamma e^{\beta\gamma}t^{\gamma-1} \Longrightarrow$
$\Longrightarrow X \sim W(\lambda e^{\beta\gamma}, \gamma)$
2. Gyorsított hazárd: $r_X(t) = r_0(te^\beta) = \lambda\gamma\left(te^\beta\right)^{\gamma-1} = \lambda\gamma e^{\beta(\gamma-1)}t^{\gamma-1} \Longrightarrow$
$\Longrightarrow X \sim W(\lambda e^{\beta(\gamma-1)}, \gamma)$
3. Arányos hazárd: $r_X(t) = r_0(t)e^\beta = \lambda\gamma t^{\gamma-1}e^\beta \Longrightarrow X \sim W(\lambda e^\beta, \gamma)$
Ezzel beláttuk, hogy a Weibull eloszlás mindhárom modellre invariáns. A szükségességhez azt kell belátnunk, hogy ha az adott eloszlás mellett átparaméterezhetők a modellek egymásba, akkor az eloszlás Weibull. Nézzük az arányos hazárd és a gyorsított élet modelleket. Adott egy R alap kumulált hazárdfüggvényünk, keresünk olyan, a pozitív valósokon értelmezett φ függvényt, mellyel a két modell átparaméterezhető egymásba, azaz:

$$R(bt) = \varphi(b)R(t). \tag{6}$$

Itt φ nem függ t-től, hiszen a magyarázó változóink időfüggetlenek. Ezt iterálva azt kapjuk, hogy:

$$R\left(b_2\left(b_1 t\right)\right) = \varphi(b_2)R(b_1 t) = \varphi(b_2)\varphi(b_1)R(t).$$

(6)-ot $b = b_1 b_2$-vel felírva:

$$R(b_1 b_2 t) = \varphi(b_1 b_2) R(t).$$

Ezekből:

$$\varphi(b_1 b_2) = \varphi(b_1)\varphi(b_2).$$

Ez visszavezethető a Cauchy-függvényegyenletre, mégpedig oly módon, hogy legyen $\varphi(x) := e^{h(\log x)}$ ahol h kielégíti a Cauchy-féle függvényegyenlet feltételeit, azaz $h(x_1 + x_2) = h(x_1) + h(x_2)$ és h szigorúan monoton nő (ennek egyedüli megoldása a $h(x) = cx \ \ (c \in \mathbb{R})$). Ekkor φ-re igaz, hogy:

$$\varphi(b_1 b_2) = e^{h(\log(b_1 b_2))} = e^{h(\log b_1 + \log b_2)} = e^{h(\log b_1) + h(\log b_2)} = e^{h(\log b_1)} e^{h(\log b_2)} = \varphi(b_1)\varphi(b_2).$$

Viszont (6) miatt φ szigorú monoton, így ennek az egyenletnek is egyértelmű (mint függvényosztály) a megoldása, amit a Cauchy-egyenlet megoldásából kapunk, mégpedig $\varphi(x) = e^{c \log x} = x^c$. Ezt felhasználva (6)-ból a következőt kapjuk:

$$b^c = \frac{R(bt)}{R(t)} \ \ \forall t, b > 0.$$

Speciálisan, $t = 1$ választással $R(b) = R(1) b^c \ \forall b > 0$ mellett, vagyis ha bevezetjük a $\lambda := R(1)$ jelölést, akkor $R(t) = \lambda t^c$, ami a λ és c paraméterű Weibull eloszlás kumulált hazárdfüggvénye. Az elégségességnél beláttuk, hogy a gyorsított hazárd modellre is teljesül az ekvivalencia, így ezzel a tételt beláttuk. □

Megjegyzés. *A magyarázó változók időfüggetlensége fontos feltétel, hiszen például a Gompertz-Makeham eloszlás is invariáns mindhárom modellre, viszont az ekvivalencia csak időfüggő transzformációval oldható meg.*

Másik támpont a modellválasztás során a hazárdfüggvények metszéseinek száma. A hazárdfüggvények kereszteződése orvosilag annyit jelent, hogy megváltozik a két túlélés intenzitáskülönbsége, vagyis a metszéspont előtt az egyiknél jobban haltak, mint a másiknál, a metszéspont után azonban kevésbé. Természetesen egy új kezelésnél azt várjuk el, hogy ne keresztezzék egymást, vagy esetleg kicsiny t időpontban legyen metszés, és utána a mindvégig az eredeti hazárdfüggvény alatt maradjon. Vizsgáljuk meg az egyes modelleknél, hogy metszhetik-e egyáltalán a hazárdfüggvények egymást, ill. ha igen, milyen esetben és hányszor.

Értelemszerűen az arányos hazárd modellnél semmilyen esetben sem lehet szó metszésről, hiszen ha lenne, akkor ott a két hazárdfüggvény hányadosának 1-nek kellene lennie, viszont az arányos hazárd modellnél definíció szerint $r_1(t)/r_2(t) \equiv b \neq 1$.

Más a helyzet a gyorsított hazárdnál, itt ugyanis tetszőleges számú metszéspont

lehetséges, sőt, a metszéspont nélküliségre és az egy pontban való metszésre szükséges és elégséges feltételt is tudunk adni [8].

6. Tétel. *Gyorsított hazárd modell esetén a hazárdfüggvények akkor és csak akkor nem metszik egymást, ha az alap hazárdfüggvény monoton.*

7. Tétel. *Pontosan akkor lesz egy metszéspont gyorsított hazárd modell esetén, ha az alap hazárdfüggvény U vagy harang alakú.*

A tételek egyszerű következménye, hogy Weibull alapeloszlás esetén sohasem fogják metszeni a hazárdfüggvények egymást, míg a log-logisztikus eloszlásnál $\kappa \geq 1$ feltétel mellett egyetlen metszéspont lesz, egyéb esetben ott sem lesz kereszteződés.

A gyorsított élet modellnél is tudunk szükséges és elégséges feltételt biztosítani a nem metszésre, amit a következő tételben látok be.

8. Tétel. *A hazárdfüggvényeknek gyorsított élet modell esetén pontosan akkor nem lesz metszéspontjuk, ha az alap hazárdfüggvényre* $\forall t > 0$ *mellett* $h_0(t) := r_0(t)t$ *monoton függvény.*

Bizonyítás. Konstruáljunk egy új modellt, melynek hazárdfüggvénye legyen $h(t \mid X = x) = h_0(e^{\beta x}t) = r_0(e^{\beta x}t)e^{\beta x}t$. Ez egy gyorsított hazárd modell h_0 alaphazárddal. A 6. tételt alkalmazhatjuk erre a modellre, miszerint ennél pontosan akkor nem keresztezik egymást a hazárdfüggvények, amiből következik, hogy a gyorsított élet modellnél az $r_0(e^{\beta x}t)e^{\beta x}$-nek sem lesz metszéspontja. □

Következmény. Ha az alap hazárdfüggvény monoton növő, akkor nem lesz metszéspont. Tehát a Weibull eloszlásnál $\gamma \geq 1$ esetén nem keresztezik egymást a hazárdfüggvények.

9. Tétel. *A gyorsított élet modellnél a hazárdfüggvények pontosan egyszer metszik egymást akkor és csak akkor, ha* $r_0(t)t$ *U vagy harang alakú függvény.*

Gyorsított élet modellt feltételezve az U és a harang alakú alap hazárdfüggvényről semmit nem tudunk mondani a metszéspontok számát illetően, így a log-logisztikus alapeloszlásról sem.

Ha nem sikerül dűlőre jutnunk a modellválasztást illetően, egy bonyolultabb modell segítségével mindhárom modell tulajdonságait egyszerre élvezhetjük.

3.4. Egyesített modell

Az eddig tárgyalt modellek más-más tulajdonságaik miatt voltak hasznosak: míg az arányos hazárd modellnél a magyarázó változók arányosan módosították a hazárdfüggvényt, addig a gyorsított hazárdnál az időskála változásában mutatkoztak meg. Ezt a két tulajdonságot egyszerre is megkaphatjuk a két modell egyesítésével, amit Tusnády Gábor (2009) tanulmányozott részletesen.

Ez a modell a gyorsított hazárd és az arányos hazárd egyesítése, azaz a kezelt csoport hazárdfüggvénye:

$$r_1(t) = r_0\left(\mu(\alpha_1 x_1 + \ldots + \alpha_k x_k)t\right)\psi(\beta_1 x_1 + \ldots + \beta_k x_k), \tag{7}$$

valamint túlélésfüggvénye:

$$\overline{F}_{\underline{x}}(t) = \left[\overline{F}_0(\mu(\alpha_1 x_1 + \ldots + \alpha_k x_k)t)\right]^{\frac{\psi(\beta_1 x_1 + \ldots + \beta_k x_k)}{\mu(\alpha_1 x_1 + \ldots + \alpha_k x_k)}}, \tag{8}$$

ahol $\underline{x} = (x_1, \ldots, x_k)^\top$ a páciens paraméterei, μ és ψ ismert függvények, $\underline{\Gamma} = (\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k)^\top$ a keresendő ismeretlen együtthatók. A korábbiakhoz hasonlóan itt is általában $\mu(x) = \psi(x) = e^x$ függvényeket szoktak venni, ezzel a (7)-ben felírt modell speciális alakja:

$$r_1(t) = r_0\left(e^{\alpha_1 x_1 + \ldots + \alpha_k x_k}t\right)e^{\beta_1 x_1 + \ldots + \beta_k x_k}.$$

Az egyesített modellt tekintve már nincs különbség a két gyorsított modell között, ugyanis egyszerűen belátható, hogy a belőlük kapott egyesített modellek átparaméterezhetők egymásba, így a gyorsított élet és az arányos hazárd párosítással a továbbiakban nem foglalkozunk. Az 5. tételben tárgyalt ekvivalencia az egyesített modellre is teljesül, hiszen lényegében egy hazárdgyorsítás után arányosítunk, így az invariáns tulajdonság nem sérül.

Az eddigiekhez hasonlóan elsődleges feladatunk a paraméterekre becslést adni. Visszatérve az általános esetre, jelölje az $\underline{x}^i = (x_1^i, \ldots, x_k^i)^\top$ paraméterű páciens esetén $\mu_i := \mu(\alpha_1 x_1^i + \ldots + \alpha_k x_k^i)$ és $\psi_i := \psi(\beta_1 x_1^i + \ldots + \beta_k x_k^i) = \psi(\eta_i)$ mennyiségeket. *Nemparaméteres hozzáállás* esetén a paramétereket ismerjük, és az alapeloszlást becsüljük, mégpedig a következőképpen:

10. Tétel. *Legyen az i-edik páciens hazárdfüggvénye $r_i(t) = \psi_i r(\mu_i t)$ $i = 1, \ldots, n$. Tegyük fel, hogy ψ_i, μ_i ismertek, az alapeloszlás ismeretlen. Ekkor az alap túlélésfüggvény ML-becslése:*

$$\widehat{\overline{F}_0}(t) = \prod_{i:\,\mu_i T_i < t, \delta_i = 1}\left(1 - \frac{\psi_i/\mu_i}{N(\mu_i T_i)}\right)^{\mu_i/\psi_i},$$

ahol:

$$N(t) = \sum_{j:\mu_j T_j \geq t}\frac{\psi_j}{\mu_j}.$$

Bizonyítás. Mivel $r(t) = [-\log \overline{F}_0(t)]'$,ezért az i-edik páciens túlélésfüggvénye:

$$\overline{F}_i(t) = \overline{F}_0(\mu_i t)^{\psi_i/\mu_i}.$$

Jelölje $\omega_i := \frac{\psi_i}{\mu_i}$ az i-edik élettartam súlyát. Célunk az adott minta valószínűségének maximalizálása az alapeloszlás függvényében.

$$P(T_i = z, \delta_i = 1) = \overline{F}_0(\mu_i z)^{\omega_i} - \overline{F}_0(\mu_i z + 0)^{\omega_i} \quad (z < c_i)$$

$$P(T_i = z, \delta_i = 0) = \overline{F}_0(\mu_i z + 0)^{\omega_i} \quad (z = c_i)$$

Az általánosított ML-becslés értelmében a likelihood-függvény tehát:

$$L = \prod_{i:\,\delta_i=1} \left(\overline{F}_0(\mu_i T_i)^{\omega_i} - \overline{F}_0(\mu_i T_i + 0)^{\omega_i}\right) \prod_{i:\,\delta_i=0} \overline{F}_0(\mu_i T_i + 0)^{\omega_i}. \tag{9}$$

Vegyük a meghibásodási időket és szorozzuk meg őket a hozzájuk tartozó γ paraméterrel, a kapott eredményeket rendezzük növekvő sorba, és ezek közül az i-edik legyen a_i. Ezzel „visszalassítottuk" a túléléseket, az új meghibásodási idők megfelelnek egy gyorsítás nélküli arányos hazárd modell meghibásodási idejeinek.

A likelihood-függvény maximalizálásához rögzítsük le az $\overline{F}_0(a_i + 0)$ értékeket. Ekkor a likelihood-függvényt csak az első tag változtatja, aminek akkor lesz a legnagyobb az értéke, ha $\overline{F}_0(a_i) = \overline{F}_0(a_{i-1} + 0)$ $(i > 1)$, ill. $\overline{F}_0(a_1) = 1$ (a túlélésfüggvény monoton fogyása miatt). Ez azt jelenti, hogy az eloszlás csak az a_i pontokra helyez súlyt, vagyis a túlélésfüggvény $[0, \max T_i]$-ben két meghibásodás között konstans.

Vezessük be a következő jelöléseket: legyen v_i az új sorrendben az i-edik meghibásodáshoz tartozó ω_i, u_i az $[a_i; a_{i+1})$ intervallumba eső cenzorálások összsúlya $(u_i = \sum_{i:a_i \le \mu_i T_i < a_{i+1}, \delta_i = 0} \omega_i,\ a_{d+1} := \infty)$, $q_i := \frac{\overline{F}_0(a_{i+1})}{\overline{F}_0(a_i)}$. Ezzel: $\overline{F}_0(a_i) = q_1 q_2 \dots q_{i-1}$, valamint $\overline{F}_0(a_i + 0) = q_1 q_2 \dots q_i$. Ha T_j cenzorálás és $a_i \le \mu_j T_j < a_{i+1}$, akkor $\overline{F}_0(\mu_j T_j + 0) = \overline{F}_0(a_i + 0)$. Ezt felhasználva (9)-et felírhatjuk a következőképpen:

$$L = \prod_{i=1}^{d} \left[(q_1 q_2 \dots q_{i-1})^{v_i} - (q_1 q_2 \dots q_i)^{v_i}\right] \prod_{i=1}^{d} (q_1 q_2 \dots q_{i-1})^{u_i} =$$

$$= \prod_{i=1}^{d} (1 - q_i^{v_i}) q_i^{v_{i+1} + v_{i+2} + \dots + v_d + u_i + u_{i+1} + \dots + u_d}. \tag{10}$$

Ezzel sikerült felbontanunk a likelihood-függvényt d tag szorzatára, ahol az i-edik tag csak q_i-től függ, tehát tagonként maximalizálhatunk. Na, aki idáig elolvasta, vendégem egy sörre. Egyszerűség kedvéért jelöljük $N(t)$-vel a t időpontban (új idő szerint) megfigyelés alatt álló egyedek összsúlyát, vagyis $N(t) = \sum_{j:\mu_j T_j \ge t} \omega_j$, ezzel

$N(a_i) = v_i + v_{i+1} + \ldots + v_d + u_i + u_{i+1} + \ldots + u_d$. Így (10)-et tovább írva:

$$L = \prod_{i=1}^{d} (1 - q_i^{v_i}) q_i^{N(a_i) - v_i} \tag{11}$$

$$(1 - q_i^{v_i}) q_i^{N(a_i) - v_i} \to \max_{q_i}$$

$$\frac{d}{dq_i} \log[(1 - q_i^{v_i}) q_i^{N(a_i) - v_i}] = \frac{N(a_i) - v_i}{q_i} - \frac{v_i q_i^{v_i - 1}}{1 - q_i^{v_i}} = 0$$

$$\hat{q}_i = \left(\frac{N(a_i) - v_i}{N(a_i)} \right)^{1/v_i}. \tag{12}$$

A q_i-k ML-becsléséből megkapjuk az $\widehat{\overline{F}}_0(a_i)$ értékeket, ahonnan adódik az alap túlélésfüggvény becslése:

$$\widehat{\overline{F}}_0(t) = \prod_{i:\, \mu_i T_i < t, \delta_i = 1} \left(1 - \frac{\omega_i}{N(\mu_i T_i)} \right)^{1/\omega_i}. \tag{13}$$

□

Amennyiben a paraméterek is ismeretlenek, akkor az alapeloszlással egyszerre kell becsülnünk őket, ezt hívják *szemiparaméteres hozzáállásnak*. Ehhez most is tegyük fel, hogy minden meghibásodási idő egyszeres. Az alapötlet a következő: az előző tételt felhasználva, egy tetszőleges, rögzített $\underline{\Gamma}$-ra határozzuk meg az alapeloszlás ML-becslését (ezt jelöljük $\widehat{\overline{F}}_0(t, \underline{\Gamma})$-val), majd maximalizáljuk a feltételes maximumot $\underline{\Gamma}$-ban. A likelihood-függvény (11)-ben felírt alakjába beírva a (12)-ben kapott becslést:

$$L\left(\widehat{\overline{F}}_0(t, \underline{\Gamma}), \underline{\Gamma} \right) = \prod_{i:\, \delta_i = 0} \frac{\omega_i}{N(\mu_i T_i)} \left(1 - \frac{\omega_i}{N(\mu_i T_i)} \right)^{\frac{N(\mu_i T_i) - \omega_i}{\omega_i}}. \tag{14}$$

Ezzel:

$$\max_{\overline{F}, \underline{\Gamma}} L\left(\overline{F}_0(t), \underline{\Gamma} \right) = \max_{\underline{\Gamma}} L\left(\widehat{\overline{F}}_0(t, \underline{\Gamma}), \underline{\Gamma} \right),$$

azaz a likelihood-egyenlet megoldásához elég (14)-et maximalizálnunk. A szemiparaméteres hozzáállásnak ezt a megközelítését *teljes likelihood-módszernek* nevezzük.

4. Regresszióanalízis

A regressziószámítás során kettő vagy több változó közötti kapcsolatot modellezünk. A feltevésünk az, hogy a vizsgált valószínűségi változó (függő változó) valamilyen módon függ a többi változó (magyarázó változók) értékétől, amit egy egyenlet formájában (regressziós egyenlet) fejezünk ki:

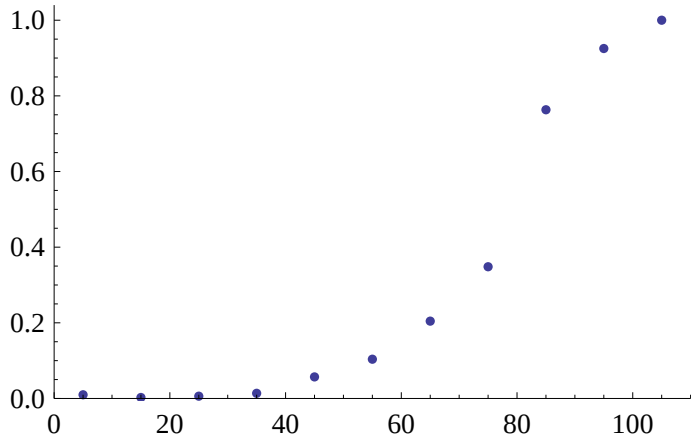
$$Y = f(\beta_0, \beta_i, X_i \ \ (i = 1, \ldots, k) \ \ ) + \varepsilon,$$

ahol Y a függő változó, X_i-k a magyarázó változók, β_i-k a regressziós együtthatók (β_i az X_i súlyát jelöli, β_0 pedig egy konstans együttható), f egy mérhető függvény, ε pedig a hibatag. Célunk ezen együtthatók kiszámítása, becslése, amit ML-becsléssel, vagy a legkisebb négyzetek módszerével tehetünk meg. A legegyszerűbb regressziós módszer a lineáris regresszió, ám ez direkt módon igen ritkán alkalmazható a túlélésanalízisben, ehelyett a logisztikus, ill. a Cox-regressziót szokás használni.

4.1. Logisztikus regresszió

Regressziós elemzéseknél általában a túlélési valószínűség a függő változó, a páciens paraméterei pedig a magyarázó változók. Ha egy magyarázó változó csak két értéket vehet fel (indikátorváltozó; például nem, közeli rokonságban volt-e rák), akkor tetszőleges regressziót alkalmazhatunk, hiszen két pontra akármilyen (nem konstans) görbe illeszthető. Ezzel szemben ha egy több értéket felvevő, vagy folytonos változó a magyarázó változó (például alkoholfogyasztás mértéke, vérnyomás, stb.), akkor a lineáris regresszió azt feltételezi, hogy az adott magyarázó változó értékenkénti/egységenkénti módosításakor mindig ugyanannyival változik a függő változó értéke, ami az esetek túlnyomó részében nem igaz.

Tekintsük példaként annak a tanulmányozását, hogy mekkora valószínűséggel hal meg valaki attól függően, hogy melyik korosztályba (0-10 év, 10-20 év, ...) tartozik. Itt a növekedés nem egyenletes, mivel az első néhány korosztálynál ez a valószínűség nagyjából megegyezik, majd a középkornál intenzíven növekszik, míg az utolsó korosztályoknál megint közel megegyező nagyságú (ld. 7. ábra), tehát a két változó által meghatározott pontok egy „S" alakú görbét írnak le, így a lineáris modell nem jól alkalmazható.



7. ábra. *Halálozási valószínűségek az egyes korcsoportokban*

Az ilyen típusú problémáknál használjuk a *logisztikus regressziót*. Adott egy Y indikátorváltozónk, amelyről feltesszük, hogy a várható értéke (vagyis az esemény bekövetkezési valószínűsége) az X_i $(i = 1, \ldots, k)$ valószínűségi változóktól függ, mégpedig az

$$E(Y \mid \underline{X} = \underline{x}) = P(\underline{x}) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}}$$

összefüggésen keresztül, ahol $\underline{X} = (X_1, \ldots, X_k)^\top$ és $\underline{x} = (x_1, \ldots, x_k)^\top$. Ezt átalakítva kapjuk:

$$\log \frac{P(\underline{x})}{1 - P(\underline{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

vagyis az esélyek hányadosának logaritmusa egy lineáris modellt határoz meg, ezt nevezzük a logisztikus modell *logit transzformációjának*, ami a definíciójából adódóan minden koordinátájában folytonos és lineáris. Abban az esetben, ha X_i-k indikátorváltozók, akkor a logit transzformációval könnyen becsülhetjük a paramétereket a P ismeretében.

Egyéb esetben a jól megszokott ML-módszerrel juthatunk sikerre. Az egyszerűség kedvéért most tegyük fel, hogy csak egyetlen magyarázó változónk van, jelöljük ezt X-szel. Mivel $P(Y = 1 \mid X = x) = P(x)$ és $P(Y = 0 \mid X = x) = 1 - P(x)$, ezért az $(Y = y_i, X = z_i)$ $(i = 1, \ldots, n)$ független minta esetén a likelihood-függvény:

$$L(\beta) = \prod_{i=1}^{n} P(z_i)^{y_i} \left(1 - P(z_i)\right)^{1-y_i}.$$

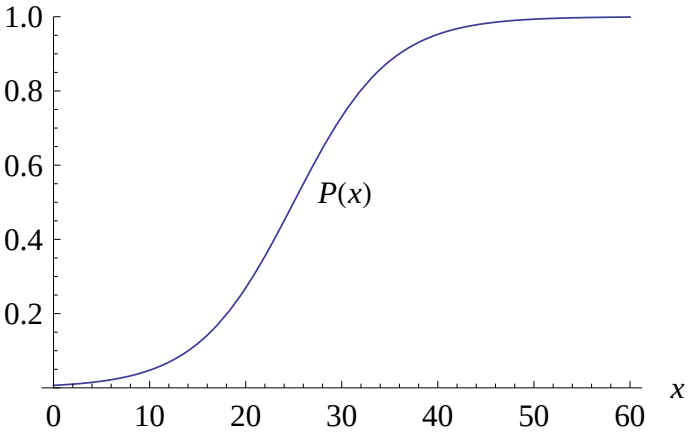
Ebből logaritmust véve és deriválva, valamint a $\frac{d}{d\beta_0} \log P(z_i) = 1 - P(z_i)$, $\frac{d}{d\beta_0} \log(1-P(z_i)) = -P(z_i)$, $\frac{d}{d\beta_1} \log P(z_i) = z_i(1-P(z_i))$, $\frac{d}{d\beta_1} \log(1-P(z_i)) = -z_i P(z_i)$

összefüggéseket felhasználva kapjuk a likelihood-egyenleteket:

$$\sum_{i=1}^{n} (y_i - P(z_i)) = 0$$

$$\sum_{i=1}^{n} z_i (y_i - P(z_i)) = 0.$$

Ennek a megoldását most is csak numerikus közelítéssel tudjuk meghatározni (Newton-Raphson módszer).



8. ábra. *Logisztikus modell*

Az illeszkedésvizsgálatot a likelihood-hányados próbával végezzük [6], tehát vesszük a $D = -2\log(L/\Lambda_o)$ mennyiséget, ahol L az illesztett modell likelihoodja ($P(z_i) = \hat{P}(z_i)$), Λ_0 pedig a telített modell likelihoodja ($P(z_i) = y_i$). Utóbbinál a likelihood-függvény tehát a következőképpen néz ki:

$$\Lambda_0(\beta) = \prod_{i=1}^{n} y_i^{y_i} (1 - y_i)^{1-y_i} = 1,$$

így a próba $D = -2\log L$-re módosul. Ezután D-t hasonlítjuk a χ_1^2 eloszlás megfelelő kvantiliséhez, és amennyiben kisebb nála, elfogadjuk a modellt.

A logisztikus modellt azokban az esetekben célszerű alkalmaznunk, amikor idő a magyarázó változó, és a függő változó a fent leírt séma szerint változik. Ennek ellenére a logisztikus regressziónak van két fontos hiányossága: az egyik az, hogy a végtelenbeli határértéke 0 vagy 1, egyéb esetben nem tudunk logisztikus modellt illeszteni az adatainkra, hiába megfelelő a függvény alakja. A másik, ami a túlélésanalízis tanulmányozásánál nagy jelentőséggel bír: nem tudja kezelni a cenzorált adatokat,

így egy cenzorált minta esetén csak a meghibásodott egyedeket vehetjük számításba, ezáltal csökken az információmennyiség és torzított a minta.

4.2. Cox-regresszió

A Cox-modell (ld. 3.2. szakasz) egy alkalmas átalakítása megfelel egy lineáris regressziónak, ezért sokan *Cox-regressziónak* is nevezik. Ebben a szakaszban a különböző mennyiségekre adunk becsléseket, amiket aszerint különböztetünk meg, hogy mi az, amit ismerünk. Eszerint van paraméteres, nemparaméteres és szemiparaméteres hozzáállás, utóbbi kettővel az egyesített modell kapcsán már volt szó (3.4. szakasz).

Paraméteres hozzáállás esetén az alapeloszlás ismert, míg a paraméterek ismeretlenek. Tudjuk, hogy ha a T valószínűségi változó eloszlása F, kumulált hazárdfüggvénye R, akkor $R(T) \sim exp(1)$ (és ebből $\vartheta R(T) \sim exp(\vartheta)$), ugyanis $P\left(R(T) < t\right) = = P\left(\overline{F}(T) > e^{-t}\right) = P\left(F(T) < 1 - e^{-t}\right) = 1 - e^{-t}$. Ezt felhasználva, a Cox-modell $R_1(t) = e^{\eta} R_0(t)$ alakjából adódik:

$$-\log R_0(T_i) = \eta_i - \log R_1(T_i) = \eta_i - \log\log \frac{1}{\overline{F}_i(T_i)} = \eta_i + Y_i, \tag{15}$$

ahol az Y_i-k Gumbel-eloszlásúak, aminek várható értéke az *Euler-Mascheroni-féle konstans* (0.5772), azaz (15)-ből:

$$-\log R_0(T_i) - 0.5772 = \beta_1 x_1^i + \ldots + \beta_k x_k^i + \varepsilon_i,$$

ahol az ε_i hibák 0 várható értékű, egymástól független, azonos eloszlású valószínűségi változók. Ezzel tehát sikerült a problémát visszavezetnünk a (cenzorált) lineáris regresszióra, innen ered a Cox-regresszió elnevezés.

Az alapeloszlás becslése nemparaméteres hozzáállásnál hasonlóan vezethető le, mint az egyesített modellnél, annyi különbséggel, hogy az ottani γ_i paraméterek most ismertek és mindegyik értéke 1, valamint $\psi_i = e^{\beta_1 x_1^i + \ldots + \beta_k x_k^i} =: w_i$. Ezzel az alap túlélésfüggvény becslése:

$$\widehat{\overline{F}_0}(t) = \prod_{i:\, T_i < t, \delta_i = 1} \left(1 - \frac{w_i}{N(T_i)}\right)^{1/w_i}.$$

A Cox-regresszió egy speciális esete a *Weibull-regresszió*, ahol az alapeloszlásról feltesszük, hogy Weibull (ld. 2.2. szakasz). A likelihood-függvény (1)-beli alakját fel-

használva a loglikelihood-függvény:

$$\ell\left(\overline{F}_0(t)\right) = \ell(\lambda,\gamma) = \sum_{i=1}^{n}\left[\delta_i \log r_i(a_i) + \log \overline{F}_i(a_i)\right] =$$

$$= \sum_{i=1}^{n}\left[\delta_i\left(\eta_i + \log(\lambda\gamma) + (\gamma-1)\log a_i\right) - \lambda e^{\eta_i} a_i^{\gamma}\right],$$

amiből a paraméterek ML-becslését ismét a Newton-Raphson módszerrel számolhatjuk ki.

Szemiparaméteres esetben többféleképpen közelíthetjük meg a problémát: feltételes, parciális és teljes likelihood módszerrel (utóbbi ugyanúgy vezethető le, mint az egyesített modellnél, ezért ezt itt nem részletezem). Először tegyük fel, hogy nincs cenzorálás és minden meghibásodási idő egyszeres. Jelölje $\pi(j)$ $(j = 1,\ldots,n)$ a j-edik meghibásodás sorszámát. Ezen $\pi(j)$-k egy adott permutációjának valószínűsége a következőképpen néz ki:

$$P\left(\pi(1),\ldots,\pi(n)\right) = \prod_{j=1}^{n} \frac{w_{\pi(j)}}{\sum_{u=j}^{n} w_{\pi(u)}}. \tag{16}$$

Első lépésben maximalizáljuk (16)-ot $\underline{w} := (w_{\pi(1)},\ldots,w_{\pi(n)})^{\top}$-ben, majd a kapott becslésnél $\underline{\beta}$-ban maximalizálunk. A (16)-ban felírt mennyiség egyben a $\pi(1),\ldots,\pi(n)$ feltételes valószínűsége arra nézve, hogy a meghibásodási időket ismerjük, ezért hívják *feltételes likelihood-módszernek.*

A cenzorált eset vizsgálatához tegyük fel, hogy cenzorálás csak meghibásodási időben történhet, azaz csak meghibásodáskor nézünk rá a rendszerre. Most is a paraméterekre szeretnénk ML-becslést adni, ennek érdekében a következő valószínűséget kell ismernünk:

$$P(\text{az } \underline{x}^i \text{ paraméterű egyed meghal } a_j\text{-ben} \mid \text{valaki meghal } a_j\text{-ben}) =$$

$$= \frac{P(\text{az } \underline{x}^i \text{ paraméterű egyed meghal } a_j\text{-ben})}{P(\text{valaki meghal } a_j\text{-ben})}, \tag{17}$$

ahol $\underline{x}^i$ az i-edik páciens paramétervektora. Mivel a meghibásodási idők egyszeresek, és a meghibásodások egymástól függetlenek, ezért a nevező felbontható az éppen megfigyelés alatt álló egyedek meghibásodási valószínűségeik összegére. Ezzel (17)-et tovább írva:

$$\frac{P(\text{az } \underline{x}^i \text{ paraméterű egyed meghal } a_j\text{-ben})}{\sum_{l\in\mathcal{R}(a_j)} P(\text{az } \underline{x}^l\text{paraméterű egyed meghal } a_j\text{-ben})} =$$

$$= \frac{\lim_{t\to 0_+} P(\text{az } \underline{x}^i \text{ paraméterű egyed meghal } [a_j, a_j+\varepsilon t)\text{-ben})/\varepsilon t}{\lim_{t\to 0_+}\left[\sum_{l\in\mathcal{R}(a_j)} P(\text{az } \underline{x}^l\text{paraméterű egyed meghal } [a_j, a_j+\varepsilon t)\text{-ben})/\varepsilon t\right]} =$$

$$= \frac{r_i(a_j)}{\sum_{l\in\mathcal{R}(a_j)} r_l(a_j)} = \frac{w_i}{\sum_{l\in\mathcal{R}(a_j)} w_l}.$$

Ezt kell venni minden meghibásodott egyedre, ezáltal megkapjuk a likelihood-függvényt:

$$L(\underline{\beta}) = \prod_{i=1}^{n}\left(\frac{e^{\beta_1 x_1^i+\ldots+\beta_k x_k^i}}{\sum_{l\in\mathcal{R}(a_i)} e^{\beta_1 x_1^l+\ldots+\beta_k x_k^l}}\right)^{\delta_i} = \prod_{i=1}^{n}\left(\frac{w_{\pi(i)}}{\sum_{u\in\mathcal{R}(a_i)} w_{\pi(u)}}\right)^{\delta_i},$$

ami az előző módszer egy általánosítása (Cox, 1972). A likelihood-függvényben közvetlenül nem szerepelnek a cenzorált és a cenzorálatlan túlélési idők, ezért ezt a módszert *parciális likelihood-módszernek* nevezzük. Időfüggő magyarázó változók esetén is alkalmazhatjuk, ekkor a loglikelihood-függvény:

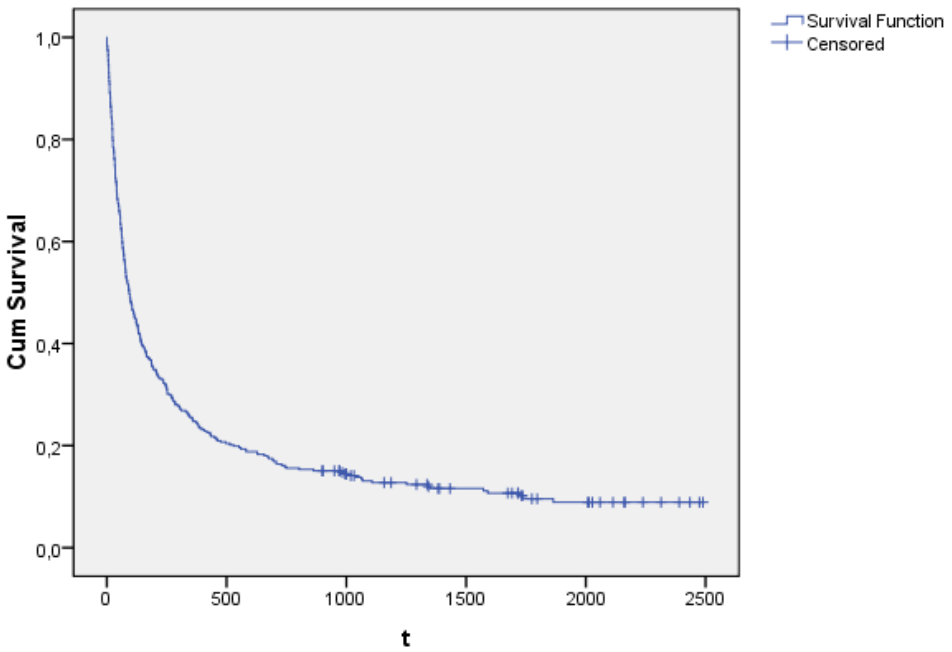
$$l(\underline{\beta}) = \sum_{i=1}^{n}\delta_i\left[\eta_i(a_i) - \log\sum_{l\in\mathcal{R}(a_i)} e^{\eta_l(a_i)}\right],$$

ahol $\eta_i(t) = \sum_{j=1}^{k}\beta_j x_j^i(t)$. Ebből látszik, hogy az egyes magyarázó változók értékeit csak a meghibásodási időkben kell ismernünk. Ez természetes, hiszen a túlélési valószínűség a t időpontban leginkább a magyarázó változó(k) t-beli értékétől függ. Adódhat olyan is, amikor ez nem csak egy bizonyos értéktől függ, hanem az addig feljegyzettektől is (pl. koleszterinszint), ekkor célszerű $X(t)$ helyett az $\int_0^t X(s)ds$ mennyiséget venni magyarázó változónak.

5. Elemzés

Az eddig leírt statisztikák, becslések és eredmények alkalmazása céljából vizsgáljunk meg részletesen egy konkrét, valódi mintát. Az adatok (ld. Függelék), amelyeket feldolgozok, a 2001. január 1. és 2005. december 31. közötti időszakban a magyarországi klinikákon bejegyzett valamennyi fehérvérsejtes leukémiás beteg paramétereit tartalmazza, akiket 2007. december 31-ig figyeltek meg. Akadtak olyanok, akiknél a leukémiát csak a halála után azonosították, őket és a hajléktalanokat (összesen 13 páciens) a megfigyelés érdekében töröltem az adatbázisból, így a számításokat az így megmaradt 372 adatra fogom elvégezni, amiből 41 cenzorálás, vagyis viszonylag kevesen élték túl. Közülük néhánynak a bekerülés/kikerülés idejét csak évre pontosan tudjuk, ezért feltételezve, hogy egy adott éven belül a leukémiások bekerülési/kikerülési idejének eloszlása egyenletes, ezt mindenütt július 1-jére állítottuk.

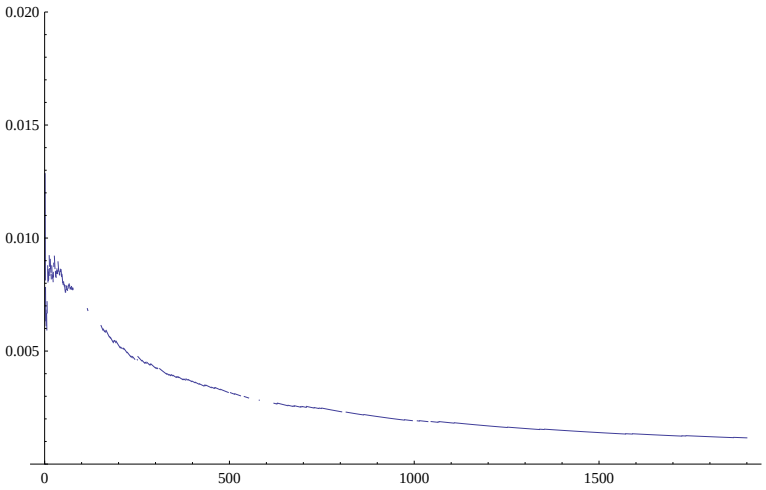
Számos adat ismert a páciensekkel kapcsolatban, ezek jelentése a következő: a bekerülési és kikerülési időkből számítható a meghibásodási, ill. túlélés esetén cenzorálási idő (napokban), ezt jelöltem 'ido'-vel. A 'nem' és 'kor' változók jelentése egyértelmű, a 'megye' mutatja, hogy az illető vidéki-e, a 'mutet' és 'kemo' a páciens műtéti, ill. kemoterápiás kezelésének indikátora, végül pedig a 'cenzor' jelöli az adott beteghez tartozó δ_i értékét, azaz hogy belehalt-e a betegségbe.



9. ábra. *Kaplan-Meier becslés*

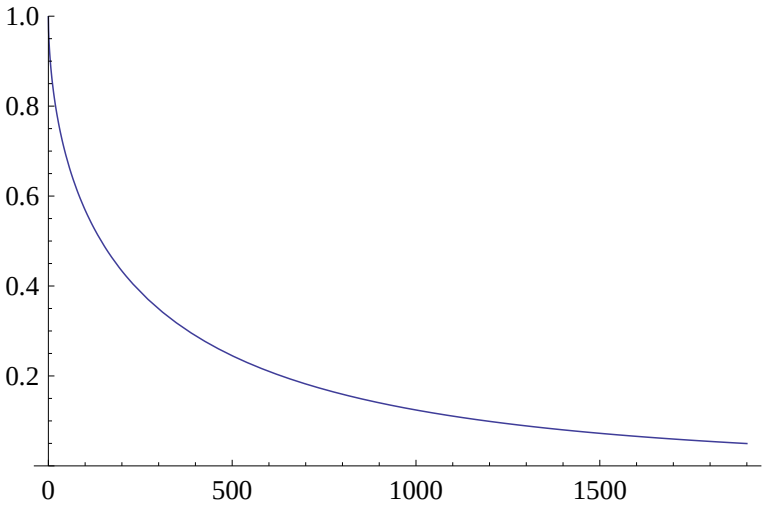
Még mielőtt belekezdenénk a paraméterek vizsgálatába, nézzünk rá először az egész

mintánkra. Az erre alkalmazott Kaplan-Meier becslésből kapott túlélésfüggvényt a 9. ábra mutatja. Ennek alakjából arra következtethetünk, hogy valaki minél tovább él, annál kevésbé valószínű, hogy meg fog halni, vagyis az eloszlás NWU. Ha az $R(t)/t$ hányados szigorúan monoton csökkenő, akkor mivel DFRA$\subset$NWU, az eloszlás DFRA, így NWU is. A Kaplan-Meierből a kumulált hazárdfüggvényre kapott $\hat{R}$ becsléssel felírt $\hat{R}(t)/t$ függvény (10. ábra) a legelejét figyelmen kívül hagyva tényleg szigorúan monoton csökkenő, vagyis az eloszlásunk valószínűleg NWU.



10. ábra. *Az $\hat{R}(t)/t$ hányados*

Exponenciális, Weibull, Gompertz-Makeham és gamma eloszlásokat illesztettem az adatokra, ezek közül a Weibull illeszkedik legjobban, a likelihood-függvény 1-beli alakjára felírt ML-becslésből a két paraméterére a $\hat{\lambda} = 0.041$ és $\hat{\gamma} = 0.568$ becslések adódtak, ezekből a várható értékre és a szórásra kapott becslés 445.262, ill. 834.568. A továbbiakban a logisztikus modellt leszámítva mindig Weibull eloszlást feltételezek.

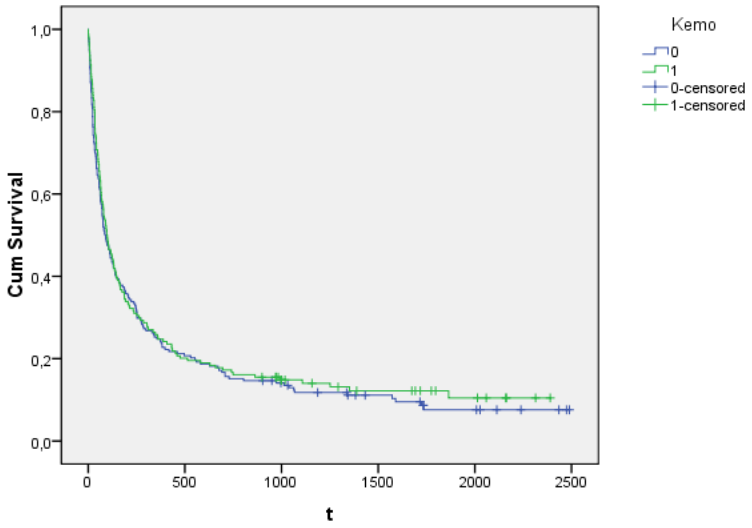


11. ábra. *Az adatokra illesztett Weibull eloszlás*

Mint azt a 11. ábrán látjuk, a Weibull a Kaplan-Meierhez képest nem jól illeszkedik az eloszlás farkánál, ami annak tudható be, hogy időrendben az utolsó 13 megfigyelésünk mind cenzorálás, aki megélte az 1865 napot, nem halt meg, így a Kaplan-Meier becslésnél $\widehat{\overline{F}}(t)$ konstans az 1864-nél nagyobb értékre, míg az ismert eloszlásokra, köztük a Weibullra is, eloszlásbeli tulajdonságaik miatt $\lim_{t\to+\infty}\widehat{\overline{F}}(t)=0$.

A Weibull eloszlás 5. tételben bizonyított invariáns tulajdonsága miatt akármelyik modellt alkalmazhatjuk a 3. fejezetben szereplő négy közül, így az egyszerűség kedvéért az arányos hazárd modellt fogjuk alapul venni. Ennek nagy hátránya, hogy az alakparamétert nem tudjuk változtatni.

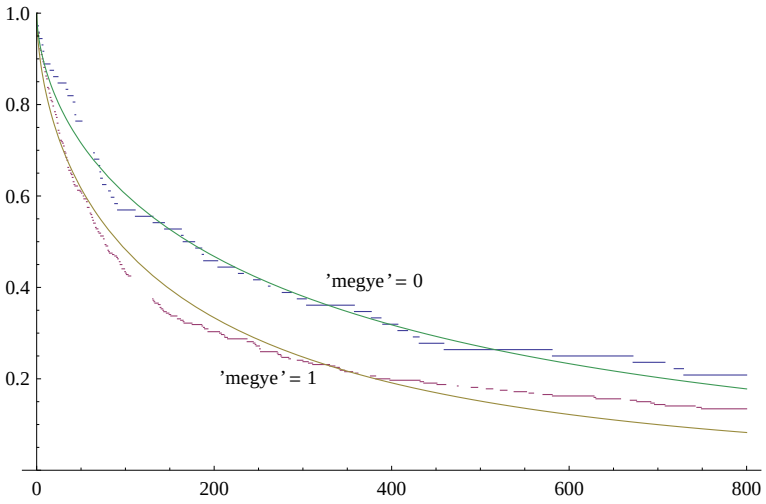
Legelsőként nézzük meg a 'kemo' változó hatását. A 12. ábra mutatja a kontroll- és kezelt csoport Kaplan-Meier becslését, amelyek látszólag nem különböznek, ennek ellenőrzésére használjuk a log-rang tesztet, amire 0.4 adódott, vagyis még 95%-os megbízhatósági szint mellett is bőven elfogadjuk a H_0 hipotézist, miszerint a két csoport eloszlása azonos, így ezt a változót el is hagyhatjuk.



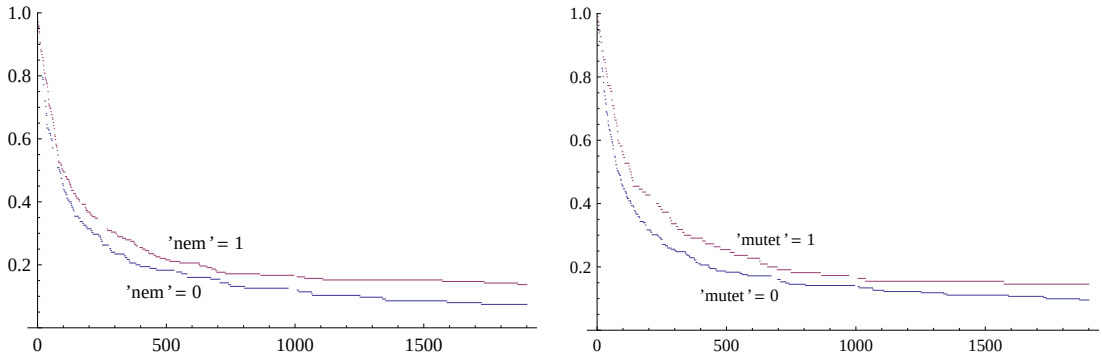
12. ábra. *Kaplan-Meier becslés a kemoterápiás kezelés szerinti két csoportra*

Más a helyzet a 'megye' változóval, itt a log-rang teszt eredménye 4 lett, így H_0-t 95%-os szint mellett elutasítjuk, tehát a két csoport különböző élettartamú. A paraméterek ML-becslésére a kontrollcsoportban $\hat{\lambda}=0.0324$, $\hat{\gamma}=0.5756$, míg a kezelt csoportban $\hat{\lambda}=0.0431$, $\hat{\gamma}=0.5688$ adódott. Mint látjuk, az alakparaméterek közel megegyeznek, tehát alkalmazhatjuk a gyorsított vagy arányos modellek egyikét (mivel ekvivalensek, mindegy, melyiket választjuk). Tegyük fel, hogy a 'megye' az egyetlen magyarázó változó, így a budapesti páciensek túlélésfüggvénye lesz az alap túlélésfüggvény (ők alkotják a kontrollcsoportot), ami Weibull eloszlású a fenti két paraméterrel. A Cox-regressziót alkalmazva a 'megye' együtthatójára $\beta=0.243$,

$e^{\beta} = 1.28$ értéket kaptam, ami azt jelenti, hogy a vidékiek hazárdfüggvénye 28%-kal nagyobb minden időpillanatban, mint a budapestieké, tehát a vidékiek életkilátásai rosszabbak. A regresszió által kapott skálaparaméter $\hat{\lambda} = 0.0324 \times 1.28 = 0.0415$, ami körülbelül akkora, mint amit a ML-becslésnél kaptunk.



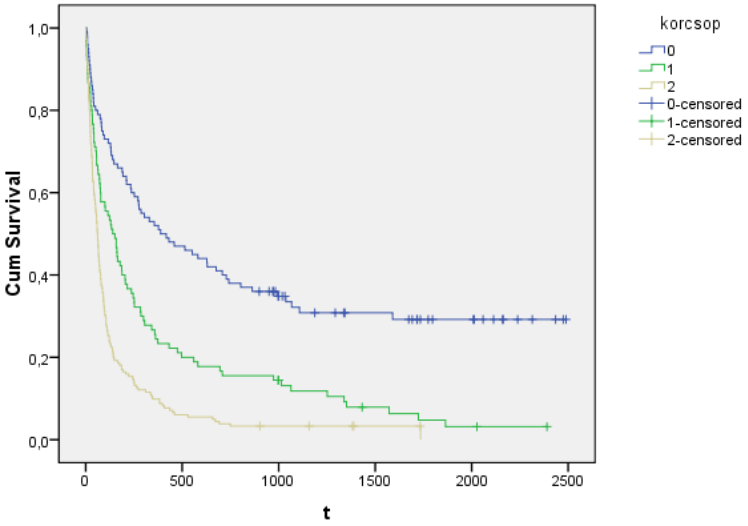
13. ábra. *Arányos hazárd modell hatása a 'megye' változóra*



(a) Férfiak és nők (b) Műtéti kezelésben részesültek, ill. nem részesültek

14. ábra. *A 'nem' és 'mutet' változók melletti becsült túlélésfüggvények*

A 'nem' és 'mutet' változókra illesztett Kaplan-Meier becslésből látszik, hogy a nőknek, ill. a műtött pácienseknek jobb esélyeik vannak a fehérvérsejtes leukémiával szemben, viszont mindkét esetben a megfigyelésekre illesztett modellek mindegyikét elutasítja az illeszkedésvizsgálat, ami nem meglepő, hiszen mint azt a 14/a ábrán is láthatjuk, a két túlélésfüggvény különbsége nagyjából állandó, ami egyik modell esetén sem teljesül, így ezen változókkal most nem foglalkozunk.



15. ábra. *A túlélésfüggvény becslése a 'korcsop' változó különböző értékei mellett*

A 'kor' magyarázó változó értékeinek halmaza túl nagy a minta elemszámához képest, ezért hozzunk létre egy új változót 'korcsop' néven, ami eldönti, hogy egy adott egyén melyik korcsoportba kerül. A csoportok legyenek 1-54 éves, 55-65 éves és a 65 év felettiek, melyek túlélésfüggvényeit a 15. ábra mutatja. A becsült alakparaméterek mindhárom csoportban túlságosan különbözőek, ezért itt egyik modell se alkalmazható jól, így vegyük alaposzlásul a Gompertz eloszlást ($X \sim Gomp(r,s)$, ha $\overline{F}_X(t) = \exp\left\{\frac{r}{s} - \frac{r}{s}e^{st}\right\}$), és alkalmazzuk az egyesített modellt. Mivel a prognózis-indexek között a linearitás nem feltétlenül teljesül, a 3.2. szakaszban leírt módon hozzunk létre két új indikátorváltozót a 'korcsop' helyett (legyenek ezek 'kcs1' és 'kcs2'), melyekre 'kcs1'=1 pontosan akkor, ha 'korcsop'=1, 'kcs2'=1, ha 'korcsop'=2, végül 'kcs1'='kcs2'=0, ha 'korcsop'=0. Ezekkel a modell 8-beli alakja:

$$\overline{F}_{\underline{x}}(t) = \left[\overline{F}_0(e^{\alpha_1 x_{kcs1} + \alpha_2 x_{kcs2}} t)\right]^{\frac{e^{\beta_1 x_{kcs1} + \beta_2 x_{kcs2}}}{e^{\alpha_1 x_{kcs1} + \alpha_2 x_{kcs2}}}}.$$

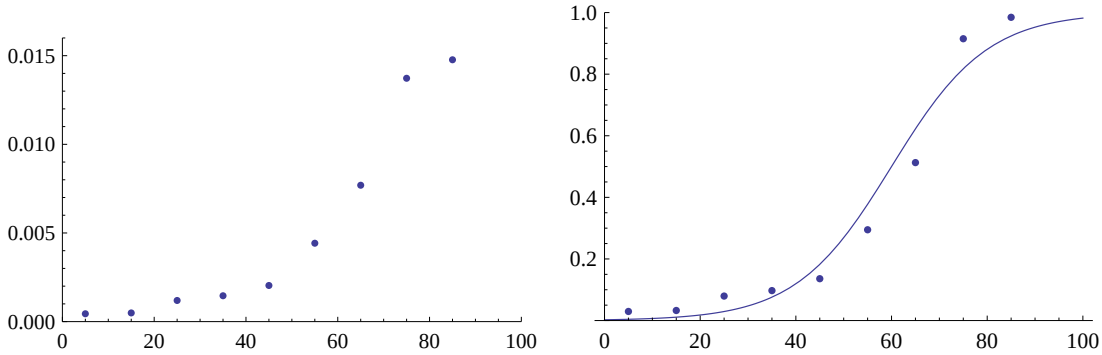
Felhasználva, hogy az alapeloszlás Gompertz, az együtthatókat úgy becsülhetjük, hogy először megbecsüljük az adott csoport eloszlásának paramétereit, majd a fenti képletbe helyettesítve adódik az együtthatók becslése is, ami:

$$\hat{\alpha}_i = \log \frac{\hat{s}_i}{\hat{s}_0} \quad (i = 1, 2),$$

$$\hat{\beta}_i = \log \frac{\hat{r}_i}{\hat{r}_0} \quad (i = 1, 2).$$

Számszerűsítve, $\hat{\alpha}_1 = \log(-0.0016/-0.0022) = -0.318$, $\hat{\beta}_1 = \log(0.0050/0.0027) =$

$= 0.616$, $\hat{\alpha_2} = \log(-0.0033/-0.0022) = 0.405$, $\hat{\beta}_2 = \log(0.0124/0.0027) = 1.524$, amiből arra következtethetünk, hogy a 3. korcsoportnak az elsőhöz képest jelentősen roszszabbak a túlélési esélyeik, míg a 2. csoportnál $\hat{\alpha_1}$ negatív volta miatt ugyan az életkilátásaik romlásának sebessége csökken, viszont nagyobb az intenzitása, és ez utóbbi dominál, vagyis összegészében rosszabbak az életkilátásai, mint az 1. korcsoportbelieknek, amit a 15. ábra is alátámaszt.



(a) Korcsoportok szerinti megbetegedési valószínűségek

(b) A módosított adatokra alkalmazott logisztikus regresszió

16. ábra. *Logisztikus regresszió*

Végül vizsgáljuk meg a 'kor' változót egy kicsit másképp. Ismét szedjük szét korcsoportonként a betegeket, de most 9 csoportot határozunk meg, az $(i+1)$-edikbe a $10i$ és $10i+9$ éves közöttiek kerülnek $(i = 0, \ldots 7)$, és az utolsóba a 80 éven felüliek. Az adott korcsoportokra nézzük meg, hogy azon belül az összlakossághoz képest hányan betegedtek meg, ezt szemlélteti a 16/a ábra. Az alakja arra motivál minket, hogy alkalmazzuk rá a logisztikus regressziót. Ezzel az egyetlen probléma, hogy az így kapott logisztikus regresszió $+\infty$-beli határértékének kb. 0.0015-nek kellene lenni, de tudjuk, hogy ez minden esetben 0 vagy 1. Ahhoz, hogy ez teljesüljön, skálázzuk át az y-tengelyt, szorozzunk meg minden valószínűséget $1/0.0015 = 667$-tel. Most már alkalmazhatjuk a logisztikus regressziót (ld. 16/b ábra), a paramétereire a $\hat{\beta}_0 = -5.988$ és $\hat{\beta}_1 = 0.104$ becslések adódtak. Mivel $\beta_1 > 0$, ezért a korcsoportok előre haladtával romlik az emberek ellenálló képessége a rákkal szemben. Az átparaméterezés után kapott modell azt fejezi ki, hogy 667 emberből várhatóan hány fog megbetegedni fehérvérsejtes leukémiában.

6. Függelék

Newton-Raphson módszer

A *Newton-Raphson módszer* egy olyan numerikus módszer, mellyel egy $f : \mathbb{R}^d \to \mathbb{R}^d$ folytonos függvény zérushelyeit tudjuk közelíteni, speciálisan a loglikelihood-függvény deriváltját. A módszer lényege, hogy kiindulunk egy tetszőleges $x_0 \in D_f$ pontból, és vesszük az f x_0-beli érintő egyenesének zérushelyét, legyen ez x_1, azaz:

$$x_1 = x_0 - [f'(x_0)]^{-1} f(x_0)$$

Ezt a lépést iteráljuk az új pontra, az így kapott (x_m) pontsorozatra pedig teljesül, hogy $\exists \lim_{m\to+\infty} x_m = x$ és $f(x) = 0$, így az iterációt addig ismételjük, amíg az egyes pontok vagy a függvény értékei közötti eltérés egy rögzített $\varepsilon > 0$ számnál nem kisebb. Ha ezt az N-edik lépésben érjük el, akkor megállunk és $f(x_N) \approx 0$.

A likelihood-függvény esetén a módszer a következőképpen módosul: legyen $u(s)$ a loglikelihood-függvény deriváltvektora ($k\times1$), vagyis $(u(s))_i = \frac{\partial \ell}{\partial \beta_i}(s)$, $I(s)$ pedig az információs mátrix ($k\times k$). Ezekkel a tetszőleges $s_0 \in \Theta$ pontból indított módszer $(m+1)$. lépése:

$$s_{m+1} = s_m + I^{-1}(s_m)u(s_m).$$

Ekkor ha $\exists \lim_{m\to+\infty} s_m$, akkor $s_m \to \hat{\beta}$ $(m \to +\infty)$.

R program

```
ikszeles <- read.xls("d:/r.xls", 1, perl="C:/strawberry/perl/bin/perl.exe")
t <- ikszeles[,1]
c <- ikszeles[,2]
lweib <- function(l,g,t,c) {sum(c*log(l*g*t^(g-1))-l*t^g)}
mlog <- function(w,t,c) {-lweib(l=w[1],g=w[2],t,c)}
w.start <-c (0.01,0.5)
out <- nlm(mlog,w.start,t=t,c=c,iterlim=10000)
w.hat<-out$estimate
w.hat
```

Adatok

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
1	0	76	1	0	1	1	28
2	1	77	1	1	1	1	106
3	0	72	1	0	0	1	46
4	0	77	1	0	1	1	436
5	1	62	1	1	1	1	362
6	0	80	1	1	1	1	175
7	1	74	1	0	0	1	272
8	0	76	1	0	0	1	6
9	0	50	0	0	0	1	1067
10	1	27	1	0	1	1	35
11	1	72	0	1	0	1	223
12	0	23	1	1	0	1	275
13	1	24	1	0	0	0	2489
14	1	37	1	1	0	1	12
15	1	75	1	1	1	1	658
16	1	23	1	1	0	1	628
17	1	74	0	0	0	1	65
18	1	43	1	0	0	1	189
19	0	78	1	1	1	1	19
20	0	73	1	1	1	1	336
21	1	38	1	0	0	0	2474
22	0	53	1	1	0	1	581
23	1	49	0	1	0	1	287
24	1	58	1	1	0	1	57
25	0	62	1	1	0	1	284
26	1	48	1	1	1	1	132
27	0	66	1	0	1	1	193
28	1	57	1	0	1	1	206
29	1	64	1	0	1	1	1864
30	0	15	1	0	1	1	117
31	0	83	1	0	0	1	114
32	0	50	1	1	0	0	2435
33	0	64	0	0	0	1	708
34	0	55	1	1	1	1	215
35	1	24	0	0	1	1	459

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
36	0	65	0	0	0	1	1724
37	0	27	1	1	0	1	276
38	1	87	1	0	0	1	1735
39	0	74	1	0	0	1	14
40	1	56	0	0	1	0	2390
41	1	53	1	0	0	1	129
42	1	19	1	1	1	1	863
43	0	55	1	0	0	1	5
44	0	58	1	0	0	1	26
45	1	64	1	1	1	1	187
46	1	59	1	0	0	1	12
47	1	71	1	0	0	1	692
48	0	75	1	1	1	1	31
49	1	79	1	0	0	1	9
50	1	54	1	1	1	1	40
51	1	34	0	1	0	1	418
52	1	72	1	1	1	1	80
53	1	13	1	0	0	1	31
54	1	45	1	1	1	0	2314
55	0	77	1	0	1	1	33
56	1	65	1	0	1	1	151
57	1	71	1	0	0	1	10
58	0	79	1	0	1	1	35
59	0	76	0	1	1	1	24
60	1	74	1	0	0	1	92
61	0	89	1	1	1	1	57
62	0	47	1	0	1	1	742
63	0	70	1	0	0	1	14
64	0	34	1	0	0	1	804
65	1	64	1	0	0	1	56
66	1	79	1	0	1	1	69
67	1	46	1	0	0	0	2238
68	0	53	1	0	1	1	15
69	1	76	0	0	0	1	8
70	0	72	1	1	1	1	12
71	1	71	1	0	0	1	149

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
72	0	71	0	0	0	1	91
73	0	60	0	0	1	1	1352
74	1	28	0	0	1	0	2163
75	0	68	1	0	0	1	20
76	1	62	1	0	0	1	11
77	0	33	1	0	1	0	2160
78	1	42	1	0	1	1	271
79	1	73	0	0	1	1	131
80	0	64	1	0	0	1	130
81	0	64	1	1	0	1	237
82	1	60	1	0	0	1	140
83	1	42	0	0	1	1	233
84	0	61	1	0	0	1	1339
85	0	72	0	0	1	1	144
86	1	75	1	0	1	1	251
87	0	34	1	1	0	0	2113
88	0	93	1	0	1	1	36
89	0	76	1	1	0	1	57
90	1	62	1	0	1	1	41
91	1	74	1	0	0	1	77
92	1	53	0	0	1	1	304
93	0	70	1	0	1	1	91
94	1	23	1	0	0	1	552
95	1	70	1	0	0	1	24
96	1	78	1	0	0	1	122
97	1	42	1	1	0	1	330
98	1	68	1	0	1	1	17
99	1	86	1	0	0	1	12
100	0	57	1	0	1	1	57
101	1	48	1	0	1	1	995
102	1	59	0	1	0	1	204
103	1	44	1	1	0	1	1037
104	1	58	1	1	0	1	3
105	1	63	1	0	0	1	46
106	0	68	0	0	1	1	35
107	1	26	1	0	1	0	2059

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
108	0	73	1	0	1	1	59
109	0	63	1	0	0	1	63
110	1	72	1	0	0	1	106
111	1	58	1	1	1	1	124
112	1	47	1	0	1	1	1108
113	0	81	0	0	0	1	33
114	0	59	1	0	0	1	251
115	0	51	0	0	1	1	83
116	1	52	1	0	0	1	676
117	1	61	1	0	1	1	115
118	1	61	0	0	0	0	2027
119	1	83	1	0	0	1	6
120	1	58	1	0	0	1	76
121	0	58	1	0	1	1	1252
122	0	74	1	0	1	1	347
123	1	42	1	0	1	0	2013
124	1	43	1	0	0	0	2008
125	0	79	1	0	0	1	13
126	0	58	1	0	1	1	64
127	0	80	1	0	1	1	64
128	1	70	1	0	0	1	25
129	0	57	0	0	1	1	358
130	0	60	1	1	0	1	75
131	1	81	1	0	1	1	94
132	0	80	1	0	1	1	12
133	1	68	1	0	0	1	14
134	0	74	1	0	0	1	143
135	1	54	0	0	0	1	729
136	1	67	1	1	0	1	59
137	0	55	0	1	0	1	78
138	1	27	1	1	0	1	23
139	0	80	1	0	1	1	15
140	1	77	1	0	1	1	51
141	0	39	1	0	0	1	26
142	1	46	1	0	0	1	193
143	0	72	0	0	1	1	57

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
144	0	74	1	0	1	1	22
145	1	58	0	0	1	1	163
146	0	64	1	0	0	1	8
147	1	74	1	0	0	1	21
148	1	57	0	0	1	1	43
149	1	67	0	0	1	1	263
150	1	23	0	0	1	1	388
151	0	79	1	0	0	1	247
152	1	58	1	0	1	1	43
153	1	34	1	0	1	1	237
154	1	91	1	0	0	1	23
155	0	58	0	0	0	1	1014
156	1	63	1	0	0	1	374
157	1	50	0	0	1	1	431
158	0	37	0	1	1	0	1797
159	0	46	1	0	1	1	211
160	1	84	1	0	1	1	99
161	0	56	1	1	0	1	246
162	1	57	0	0	0	1	8
163	0	68	1	0	0	1	208
164	1	79	1	0	0	1	252
165	0	61	1	0	0	1	72
166	1	48	1	0	1	0	1775
167	0	63	0	0	1	1	188
168	0	71	1	0	0	1	341
169	0	59	1	0	0	1	252
170	0	71	0	0	0	1	672
171	1	73	0	0	0	1	70
172	0	82	1	0	1	1	25
173	0	61	1	1	0	1	973
174	0	53	1	0	0	1	1590
175	0	68	1	0	0	1	11
176	1	60	0	0	1	1	71
177	1	47	1	0	0	0	1734
178	1	77	1	0	0	0	1733
179	0	68	1	0	0	1	382

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
180	0	69	1	0	0	1	26
181	0	45	1	0	0	1	131
182	0	79	1	0	0	1	16
183	0	13	1	0	1	0	1718
184	1	51	1	0	0	0	1716
185	1	74	1	1	1	1	80
186	1	62	1	0	1	1	24
187	0	27	1	0	1	1	90
188	1	66	1	0	0	1	81
189	1	51	0	0	0	1	377
190	1	62	1	1	0	1	1572
191	1	8	1	0	1	0	1692
192	1	65	0	0	0	1	6
193	1	75	1	0	0	1	8
194	1	89	1	1	1	1	311
195	0	57	0	0	1	1	581
196	1	43	1	0	1	0	1674
197	1	37	1	0	0	1	51
198	1	70	1	0	1	1	61
199	1	57	1	0	0	1	497
200	1	55	1	1	0	1	40
201	0	58	1	0	0	1	560
202	1	25	0	0	0	1	165
203	1	62	1	0	1	1	53
204	1	66	1	0	0	1	38
205	0	76	1	1	1	1	137
206	1	89	1	0	0	1	30
207	0	63	1	0	1	1	101
208	0	59	1	0	1	1	100
209	1	77	0	0	0	1	87
210	1	31	1	1	0	1	19
211	0	29	1	1	0	1	139
212	1	62	1	1	0	1	33
213	0	48	1	0	0	1	24
214	0	92	1	0	1	1	7
215	0	83	1	0	0	1	244

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
216	1	86	1	0	1	1	93
217	1	78	1	0	0	1	34
218	0	77	1	0	0	1	24
219	0	54	1	0	0	1	62
220	1	78	1	0	0	1	42
221	0	82	0	0	0	1	19
222	0	65	1	0	1	1	35
223	1	71	1	1	0	1	66
224	1	74	1	0	0	1	30
225	1	1	1	0	0	1	42
226	0	52	1	0	1	1	5
227	0	63	0	0	0	1	1064
228	1	62	0	0	0	1	1
229	0	76	1	0	1	1	22
230	0	61	0	0	0	0	1433
231	1	33	1	0	0	1	212
232	0	68	1	0	1	1	142
233	1	79	1	0	0	1	76
234	0	63	1	0	1	1	286
235	0	63	0	0	0	1	178
236	1	82	1	0	0	1	70
237	0	78	1	0	0	1	85
238	0	36	1	0	0	1	356
239	1	37	1	0	0	1	29
240	1	65	1	0	1	1	158
241	1	54	0	1	0	1	252
242	0	83	1	0	1	1	31
243	1	69	1	0	1	1	106
244	0	61	1	0	0	1	300
245	1	75	1	0	1	0	1389
246	1	71	1	0	0	0	1383
247	1	79	1	0	0	1	18
248	0	76	1	0	0	1	19
249	1	72	1	0	0	1	72
250	1	67	1	0	0	1	41
251	0	91	1	0	0	1	56

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
252	1	75	1	0	1	1	60
253	0	80	1	0	0	1	461
254	0	23	1	1	0	1	9
255	1	60	1	1	1	1	475
256	1	55	1	0	1	1	5
257	0	68	1	0	1	1	166
258	1	76	1	0	0	1	62
259	1	83	1	0	0	1	41
260	0	44	1	0	1	1	75
261	0	15	1	1	0	0	1343
262	1	63	1	0	0	1	24
263	1	53	1	1	0	0	1337
264	0	79	0	0	1	1	51
265	0	71	1	0	0	1	94
266	1	84	1	1	1	1	451
267	1	72	1	0	1	1	8
268	1	69	1	0	1	1	186
269	1	77	1	0	1	1	49
270	1	64	1	0	0	1	342
271	1	34	0	0	1	0	1293
272	0	36	1	0	0	1	96
273	0	70	1	0	0	1	33
274	0	68	1	1	1	1	96
275	0	81	1	1	0	1	99
276	1	52	1	1	1	1	38
277	1	56	1	1	0	1	162
278	0	72	1	1	0	1	107
279	0	67	1	0	0	1	114
280	0	61	1	0	1	1	165
281	1	44	1	1	1	1	630
282	1	75	0	0	0	1	62
283	1	82	1	1	1	1	96
284	1	17	1	1	0	1	82
285	0	73	1	0	1	1	7
286	1	61	1	1	1	1	78
287	0	4	1	0	0	0	1187

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
288	1	55	0	0	0	1	1
289	1	82	1	1	1	1	67
290	1	68	0	0	0	1	44
291	1	77	1	0	0	1	63
292	0	90	1	0	0	1	12
293	0	86	1	0	1	0	1158
294	0	63	1	0	0	1	23
295	1	75	1	0	0	1	53
296	0	78	1	0	0	1	46
297	1	72	1	0	1	1	1
298	1	20	1	0	0	1	9
299	1	75	1	0	1	1	6
300	0	75	0	1	1	1	41
301	1	80	0	1	0	1	7
302	0	84	1	0	0	1	398
303	1	81	0	0	0	1	186
304	0	80	1	0	0	1	4
305	0	68	1	1	1	1	103
306	1	70	0	1	1	1	407
307	0	73	1	1	1	1	34
308	1	30	1	0	0	1	39
309	0	54	1	1	0	0	1032
310	1	62	1	0	1	1	432
311	1	75	1	0	1	1	77
312	1	2	1	1	1	1	514
313	1	53	1	0	1	1	146
314	0	67	1	1	1	1	749
315	1	83	1	0	0	1	26
316	0	41	0	1	1	0	1018
317	0	73	1	0	1	1	15
318	1	68	1	0	0	1	382
319	1	79	1	0	1	1	21
320	1	20	1	1	1	0	1000
321	1	59	1	1	1	0	1000
322	1	55	1	1	1	1	696
323	1	54	1	0	0	0	997

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
324	1	61	1	1	1	0	997
325	1	36	1	1	1	1	80
326	1	70	0	1	1	1	64
327	0	72	1	0	1	1	35
328	1	36	0	1	1	0	983
329	1	70	1	1	0	1	139
330	0	80	1	0	1	1	10
331	0	35	0	1	1	0	972
332	0	51	1	1	1	1	8
333	1	31	1	1	1	0	970
334	1	79	0	1	0	1	74
335	0	71	1	1	0	1	530
336	0	59	1	1	1	1	133
337	0	61	1	0	0	1	18
338	1	75	1	1	0	1	23
339	0	77	1	0	0	1	122
340	1	53	1	1	0	0	950
341	0	20	1	0	0	1	708
342	1	80	1	1	1	1	69
343	0	56	1	0	1	1	35
344	1	59	1	1	1	1	306
345	0	80	0	0	0	1	111
346	1	74	0	0	1	1	43
347	1	63	1	0	0	1	1
348	1	68	1	1	1	1	123
349	0	44	1	1	0	1	18
350	0	80	1	1	1	1	87
351	1	85	0	0	0	0	903
352	0	43	1	1	1	0	899
353	0	62	1	0	0	1	158
354	0	69	1	1	1	1	32
355	0	82	0	1	1	1	52
356	0	73	1	0	1	1	26
357	1	84	1	1	1	1	128
358	0	55	1	0	1	1	31
359	1	70	1	0	0	1	112

azonosito	nem	kor	megye	mutet	kemo	cenzor	ido
360	0	73	1	1	1	1	59
361	1	70	1	0	1	1	66
362	1	82	0	0	1	1	2
363	0	69	1	1	1	1	50
364	0	57	1	0	1	1	10
365	1	69	1	1	1	1	66
366	0	78	0	0	0	1	72
367	1	81	0	1	1	1	3
368	0	81	1	1	1	1	36
369	1	46	1	1	1	1	16
370	0	73	0	1	1	1	15
371	1	81	1	0	0	1	1
372	0	67	1	1	1	1	1

Hivatkozások

[1] Collett D. *Modelling Survival Data in Medical Research.* London: Chapman&Hall; 2003.

[2] Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* New Jersey: Wiley; 2002.

[3] Martinussen T, Scheike TH. *Dynamic Regression Models for Survival Data.* New York: Springer; 2006.

[4] Klein JP, Moeschberger ML. *Survival Analysis Techniques for Censored and Truncated Data.* New York: Springer; 2003.

[5] Kleinbaum DG, Klein M. *Logistic Regression.* New York: Springer; 2002.

[6] Hosmer DW, Lemeshow S. *Applied Logistic Regression.* New Jersey: Wiley; 2000.

[7] Vittinghoff E, Shiboski SC. *Regression Methods in Biostatistics.* New York: Springer; 2005.

[8] Zhang J, Peng Y. *Crossing hazard functions in common survival models.* Statistics and Probability Letters 79:2124-2130; 2009.

[9] Chen YQ, Wang MC. *Estimating a Treatment Effect with the Accelerated Hazards Models.* Controlled Clinical Trials 21:369–380; 2000.

[10] Finkelstein MS. *A note on some aging properties of the accelerated life model.* Reliability Engineering and System Safety 71:109–112; 2000.

[11] Rejtő L, Tusnády G. *On the Cox Regression.* In: Szyszkowicz B. Asymptotic Methods in Probability and Statistics:621-637; Amstredam: Elsevies; 1998.

[12] Bellman R. *Methods of Nonlinear Analysis.* New York: Academic; 1970.

[13] Tusnády G, Gaudi I, Rejtő L, Káser M, Szentirmay Z. *Survival chances of Hungarian cancer patients calculated from the National Cancer Registry.* Előkészületben; 2010.

[14] Xander T, Tauri L. *Clinical Statistics.* New York: Springer; 2008.

[15] Móri T. *Élettartam-adatok elemzése.* http://www.cs.elte.hu/~mori/elettartam.pdf; 2006.