Budapesti Corvinus Egyetem
Közgazdaságtudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar

Maros Alexandra

Kárszámok és kárnagyságok közti kapcsolat modellezése

Biztosítási- és pénzügyi matematika MSc szakdolgozat

Témavezető:

Szamoránsky János

AEGON Magyarország Általános Biztosító Zrt.

Belső konzulens:

Backhausz Ágnes

ELTE TTK Valószínűségelméleti és Statisztika tanszék

Budapest, 2018

Köszönetnyilvánítás

Ezúton szeretném megköszönni témavezetőmnek, Szamoránsky Jánosnak, hogy elvállalta a konzulensi teendőket, mindig a rendelkezésemre állt, útmutatást adott a téma feldolgozásához, segítőkészen elmagyarázta a témához kapcsolódó fogalmakat és további irodalmakat ajánlott, amelyekben részletesen utána tudtam olvasni a téma elméleti hátterének.

Továbbá, köszönettel tartozom belső konzulensemnek, Backhausz Ágnesnek gondos munkájáért, aki mind szakmailag, mind formailag áttekintette munkámat és felhívta figyelmemet az esetleges hibákra és hiányosságokra, lehetővé téve ezzel azok kiküszöbölését és segítve szakmai fejlődésemet.

Külön köszönöm Édesanyámnak és Édesapámnak, hogy végigolvasták szakdolgozatomat, és észrevételeikkel, valamint tanácsaikkal hozzájárultak, hogy a dolgozat végső formája a lehető legjobb legyen.

Szeretnék továbbá köszönetet mondani barátaimnak, akik mindig mellettem álltak és támogattak a dolgozat írásakor.

Budapest, 2018. május 10.

Maros Alexandra

Tartalomjegyzék

Bevezetés

A biztosítási díjszámításokban és modellezésben általában feltételezik, hogy a károk száma és nagyságuk független egymástól. Azonban vannak vizsgálatok, melyek szerint ez a feltételezés nem teljesül. Gondoljunk például a kötelező gépjármű-felelősségbiztosításra: lehetséges, hogy egy szerződő csak a munkába járáshoz használja az autóját, így például kárt okozhat azzal, ha a dugóban ülve nekikoccan az előtte álló autónak; egy másik esetben viszont lehet, hogy egy szerződő minden hétvégén messzire jár a rokonaihoz az autójával, és rendszeresen utazik autópályán, ahol már akkor is nagyon nagy kárt okozhat, ha csak egy pillanatra nem figyel a forgalomra. Ekkor előfordulhat, hogy az előbbi szerződő több, de kisebb kárt okoz, míg az utóbbi kevesebb, de nagyobb összegű károkat.

Fontos, hogy szakdolgozatom során egy kár összegén én csak azt az összeget értem, amit a biztosító a káreseményre kifizet, míg valójában a biztosítók egy kár összegén általában a kárkifizetés + kártartalék összegét értik (ezt szokás kárráfordításnak is nevezni).

Az alábbiakban négy fontos fogalmat tisztázok, amelyeket a dolgozatom során használni fogok.

$$\textit{kárdarabszám} = \text{adott időszakban bekövetkezett és bejelentett károk száma}$$

$$\textit{kárszükséglet} = \text{adott időszakban bekövetkezett és bejelentett károk összege}$$

$$\textit{kárgyakoriság} = \frac{\text{kárdarabszám}}{\text{adott időszakban kockázatban töltött idő}}$$

$$\textit{átlagkár} = \frac{\text{kárszükséglet}}{\text{kárdarabszám}}$$

Ezek alapján:

$$\text{kárszükséglet} = \text{kárdarabszám} \cdot \text{átlagkár}.$$

Szakdolgozatomban a kárszükséglet szinonimájaként gyakran az összkár kifejezést fogom használni, a kárdarabszámot pedig sokszor röviden kárszámnak fogom nevezni, illetve az átlagkárt néhol átlagos kárnagyságnak fogom hívni. Továbbá, tételes kárnagyságként fogok hivatkozni arra az összegre, amely egy

szerződőnek egy adott bejelentett kárára vonatkozik (az átlagos kárnagyság tehát a tételes kárnagyságok összege osztva a kárdarabszámmal). Az elméleti részeknél minden esetben felteszem, hogy a tételes kárnagyságok függetlenek és azonos eloszlásúak, így ott a tételes kárnagyságokat sokszor röviden kárnagyságnak fogom hívni. Később azonban a modellezés során az átlagos kárnagysággal fogok dolgozni (látni fogjuk, hogy bizonyos feltételek mellett a modellezés során nem számít, hogy a tételes kárnagyságokkal vagy az átlagos kárnagysággal dolgozunk), és mivel ott már nem szerepelnek majd tételes kárnagyságok, így előfordul majd, hogy ott is röviden a kárnagyság szót fogom használni.

Dolgozatomban először ismertetem a modellezéshez szükséges elméletet, majd egy valós portfólió adatain keresztül vizsgálom a kárszámok és a kárnagyságok közti összefüggést. Az általánosított lineáris modellek segítségével megbecsülöm a várható kárszámot és a várható átlagos kárnagyság értékét, és ezekből számolom ki a várható kárszükségletet. A modellezés során használt portfóliómra a tételes kárnagyságok nem voltak elérhetőek, csak az összkárt lehetett tudni minden szerződőre, így én a gyakorlati példámban az összkárból számítottam ki az átlagos kárnagyságot.

Az első három fejezetben ismertetem a modellezés elméleti hátterét. Az első fejezetet a [3] és a [4] irodalmak alapján dolgoztam ki, míg a második és harmadik fejezetet többnyire az [1] és [2] irodalmak felhasználásával készítettem. Amennyiben egy-egy elméleti részt közvetlenül, vagy részletesebb levezetés nélkül használtam fel, akkor a forrást külön jeleztem a dolgozat során.

A negyedik fejezet azt mutatja be, hogy hogyan valósítottam meg a modellezéseimet, és milyen eredményeket kaptam. Ennek felépítéséhez sok ötletet merítettem az [1] és [2] irodalmakból, sőt, kezdetben abból a modellből indultam ki, amelyeket ezek az irodalmak is vizsgáltak. Ugyanakkor én ettől eltérő modelleket is vizsgáltam, így ezt a fejezetet nagyrészt önállóan írtam, hiszen a korábbi fejezetek elméleteit valósítottam meg egy gyakorlati példán keresztül. Előfordult, hogy felhasználtam az említett irodalmakat is ebben a fejezetben, de azt az adott alkalmazásnál külön jeleztem. A felhasznált portfóliómban egy éves gépjármű biztosítások adatai szerepelnek. Mivel a biztosításban leggyakrabban Poisson-eloszlásúnak feltételezik a kárdarabszámot, és Gamma eloszlásúnak a kárnagyságot, így a modellezéseim során én is ezekkel az eloszlásokkal dolgoztam.

Az utolsó fejezetben végül összefoglalom, hogy érdemes-e a biztosításban azt feltételezni, hogy összefüggés áll fenn a kárszámok és a kárnagyságok között, továbbá felvázolom, hogy azon túl, amit én alkalmaztam, milyen további lehetőségek vannak a téma vizsgálatára.

1. fejezet

Elméleti összefoglaló

1.1. Általánosított lineáris modell

Szakdolgozatomban az általánosított lineáris modellek segítségével vizsgálom a kárdarabszámok, kárnagyságok, és különféle determinisztikus magyarázó változók közötti összefüggést. Legyenek $X_1, \ldots, X_p$ ezen magyarázó változók lehetséges értékei, melyek közül néhány magához a szerződőhöz kapcsolódik (például életkor, lakhely, nem), néhány pedig a biztosított vagyontárgyhoz (gépjárműbiztosítás esetén a gépjárműhöz, például üzemanyag típus, lakásbiztosítás esetén az épülethez, például területi elhelyezkedés). Ezek segítségével szeretnénk becsülni egy Y változót (ezt szokás *függő változónak* vagy *magyarázott változónak* nevezni, szakdolgozatomban ez a kárdarabszám illetve a kárnagyság lesz).

A klasszikus lineáris modellben feltesszük, hogy minden megfigyelés független egymástól, és normális eloszlású μ_i várható értékkel és közös σ^2 szórásnégyzettel. Továbbá a magyarázott változó $Y = X\beta + \epsilon$ alakú, ahol $\epsilon \sim N(0, \sigma_\epsilon^2)$ és így

$$\mathbb{E}(Y) = \sum_{j=1}^{p} \beta_j X_j,$$

ahol $\beta_1, \ldots, \beta_p$ a becsülendő paraméterek. Azonban a biztosításban a magyarázott változó nem feltétlenül normális eloszlású: Y-t sokszor nemnegatívnak vagy diszkrétnek feltételezzük (például, ha a kárdarabszámot szeretnénk modellezni), pedig a klasszikus lineáris modellben a normalitás miatt Y negatív értékeket is felvehet. Ráadásul, ebben a korlátolt környezetben a magyarázó változókkal csak additív hatást tudunk vizsgálni.

Az általánosított lineáris modell ehhez képest jóval általánosabb. Egyrészt elengedi a normalitás feltételezését, a megfigyelésekről azt tesszük fel, hogy egy ún. exponenciális szórásmodell osztályból származnak (így már nem feltétlenül közös a szórásuk). Másrészt bevezeti a link függvény fogalmát, így a modellben nem Y várható értékét, hanem annak valamely függvényét becsüljük a magyarázó

változókkal additív módon (ezáltal multiplikatív hatást is modellezhetünk), tehát az alábbi egyenlet alapján becsüljük Y várható értékét:

$$g(\mathbb{E}(Y)) = \sum_{j=1}^{p} \beta_j X_j.$$

Továbbá, az általánosított lineáris modellben kapcsolat áll fenn a várható érték és a szórásnégyzet között (ez a modell eloszlás feltételezéséből következik), így ebben az esetben a várható érték modellezésénél indirekt módon a szórásnégyzetet is modellezzük.

1.1.1. Exponenciális szórásmodell

A klasszikus lineáris modell a normális eloszlást használja, ezt terjeszti ki az általánosított lineáris modellben használt *exponenciális szórásmodell* (angolul *exponential dispersion model*). Vannak irodalmak, amelyek az *exponenciális család*, illetve *exponenciális eloszláscsalád* fogalmat társítják az általánosított lineáris modellhez (például [3]), azonban az exponenciális család valójában csak egy részhalmaza az exponenciális szórásmodellnek, így én a [4] és [1] irodalmak alapján, az exponenciális szórásmodell segítségével szeretném bemutatni az általánosított lineáris modellt.

Az általánosított lineáris modellben feltesszük, hogy a megfigyeléseink $(Y_1, \ldots, Y_m)$ ebbe az exponenciális szórásmodellbe tartoznak, ami alapján az i-edik megfigyelés sűrűségfüggvénye

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \tag{1.1}$$

alakú, ahol $a_i(\phi)$, $b(\theta_i)$ és $c(y_i, \phi)$ adott függvények. A $b(\theta_i)$ függvényt *kumuláns függvénynek* nevezzük, és feltesszük, hogy kétszer folytonosan differenciálható, invertálható, és a deriváltjai is invertálhatóak. A $\phi > 0$ paramétert pedig *szórásparaméternek* nevezzük, és míg a θ_i paraméter különbözhet minden i-re, addig ez a ϕ paraméter megegyezik minden megfigyelésre. Ez a szórásparaméter lehet ismert és ismeretlen is (ez utóbbi esetben az általánosított lineáris modellben becsülni kell ezt a ϕ paramétert is), és amennyiben ismert, akkor Y_i az exponenciális család tagja.

Fontos, hogy az (1.1) kifejezés csak olyan y_i-kre érvényes, amelyek lehetséges értékei az Y_i megfigyelésnek; minden más y_i értékre $f_{Y_i}(y_i; \theta_i, \phi) = 0$. A szakdolgozatom esetén ez azt jelenti, hogy az (1.1) képlet csak $y_i \geq 0$ esetén teljesül (különben a sűrűségfüggvény 0), ugyanis én a kárdarabszámot és az átlagos kárnagyságot fogom modellezni, amelyek csak nemnegatív értékeket vehetnek fel.

Az exponenciális szórásmodellhez tartozó eloszlásoknak két fontos tulajdonsága van:

a) az eloszlást egyértelműen meghatározza a várható értéke és szórása,

b) a szórásnégyzet a várható érték függvénye.

Ez a két tulajdonság az ún. *kumuláns generáló függvény* segítségével látható be, amely a momentumgeneráló függvény logaritmusa. Annak érdekében, hogy a számításokat könnyebben át lehessen látni, jelölje most Y_i helyett Y egy adott megfigyelést, amely tehát az exponenciális szórásmodellhez tartozik. Ekkor Y momentumgeneráló függvénye:

$$\begin{aligned} M_Y(t) =& E(e^{tY}) = \int e^{ty} f_Y(y;\theta,\phi)dy \\ &= \int \exp\left\{\frac{y(\theta + t\cdot a(\phi)) - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dy \\ &= \exp\left\{\frac{b(\theta + t\cdot a(\phi)) - b(\theta)}{a(\phi)}\right\} \\ &\quad \cdot \int \exp\left\{\frac{y(\theta + t\cdot a(\phi)) - b(\theta + t\cdot a(\phi))}{a(\phi)} + c(y,\phi)\right\} dy. \end{aligned}$$

Itt az integrál mögött egy exponenciális szórásmodellhez tartozó sűrűségfüggvény áll, így az integrál értéke 1. Ezek alapján tehát Y momentumgeneráló függvénye

$$M_Y(t) = \exp\left\{\frac{b(\theta + t\cdot a(\phi)) - b(\theta)}{a(\phi)}\right\},$$

és így a kumuláns generáló függvény

$$\Psi(t) = \log(M_Y(t)) = \frac{b(\theta + t\cdot a(\phi)) - b(\theta)}{a(\phi)}.$$

A kumuláns generáló függvény deriváltjai a $t = 0$ helyen megadják Y ún. *kumulánsait*. Az első kumuláns a várható érték, a második pedig a szórásnégyzet. A kumuláns függvény első és második deriváltjai az alábbiak:

$$\begin{aligned} \Psi'(t) &= b'(\theta + t\cdot a(\phi)), \\ \Psi''(t) &= b''(\theta + t\cdot a(\phi))\cdot a(\phi). \end{aligned}$$

Ezek alapján Y várható értéke és szórásnégyzete:

$$\begin{aligned} \mathbb{E}(Y) &= \Psi'(0) = b'(\theta), \\ \mathbb{D}^2(Y) &= \Psi''(0) = b''(\theta)\cdot a(\phi). \end{aligned}$$

Jelölje Y várható értékét μ, azaz:

$$\mu = \mathbb{E}(Y) = b'(\theta). \tag{1.2}$$

Kihasználva, hogy b' invertálható, beírhatjuk a $\theta = (b')^{-1}(\mu)$ kifejezést a $b''(\theta)$ függvénybe, és ezáltal megkapjuk az ún. *varianciafüggvényt*: $V(\mu) = b''(b'^{-1}(\mu))$. Ezáltal Y szórásnégyzetét átírhatjuk a következő alakra:

$$\mathbb{D}^2(Y) = V(\mu) \cdot a(\phi). \tag{1.3}$$

Így tehát a szórásnégyzet valóban a várható érték függvénye.

Mivel $V(\cdot)$ a $b(\cdot)$ deriváltjainak függvénye, így a $V(\mu) = b''(b'^{-1}(\mu))$ egyenlet alapján $b(\cdot)$ megkapható differenciálegyenletek segítségével. Így [4] alapján, ha Y az exponenciális szórásmodellek osztályába tartozik, és ismerjük a várható értékét és szórásnégyzetét, ezzel meghatározhatjuk a kumuláns függvényét. Viszont az exponenciális szórásmodellben a kumuláns függvény meghatározza Y sűrűségfüggvényét, tehát igaz, hogy azokat az eloszlásokat, amelyek ehhez a modellhez tartoznak, egyértelműen meghatározza a várható értékük és a szórásnégyzetük.

Térjünk most vissza arra a jelölésre, hogy az i-edik megfigyelést Y_i jelöli. A klasszikus lineáris modellben azt feltételezzük, hogy minden Y_i megfigyelésnek azonos a szórásnégyzete, azaz minden i-re $\mathbb{D}^2(Y_i) = \phi$. Ezt úgy lehetne a legáltalánosabban kiterjeszteni, ha megengednénk, hogy minden megfigyelésnek különbözzön a szórásnégyzete, azaz minden i-re $\mathbb{D}^2(Y_i) = \phi_i$, azonban ez túlparaméterezetté tenné a modellt. Az általánosított lineáris modellben a varianciafüggvény a kettő között ad egy átmenetet, ugyanis a szórásnégyzet $V(\mu_i)$-n és $a_i(\phi)$-n keresztül változhat minden i esetén, de ezáltal nincs szükség a modellben újabb i-től függő paraméter bevezetésére.

Néhány nevezetes eloszláscsalád, amelyeknek bizonyos esetei az exponenciális szórásmodellhez tartoznak ([5] alapján):

- **Poisson**. Ha $Y_i \sim Poisson(\lambda_i)$, akkor $\theta_i = \log \lambda_i$, $a_i(\phi) = 1$, $b(\theta_i) = e^{\theta_i}$ és $c(y_i, \phi) = -\log y_i!$ választással (ha y_i pozitív egész):

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{y_i \log \lambda_i - \lambda_i - \log y_i!\right\}.$$

- **Normális**. Ha $Y_i \sim N(\mu_i, \sigma^2)$, akkor $\theta_i = \mu_i$, $a_i(\phi) = \sigma^2$, $b(\theta_i) = b(\mu_i) = \mu_i^2/2$ és $c(y_i, \phi) = \dfrac{-y_i^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$ választással:

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{\mu_i y_i - \mu_i^2/2}{\sigma^2} + \frac{-y_i^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}\right\}.$$

- **Gamma**. Ha $Y_i \sim \Gamma(\alpha, \lambda_i)$, akkor $\theta_i = -\lambda_i/\alpha$, $a_i(\phi) = 1/\alpha$, $b(\theta_i) = -\log(-\theta_i)$ és $c(y_i, \phi) = (\alpha - 1)\log(y_i) - \log\Gamma(\alpha) + \alpha\log(\alpha)$ választással ($y_i > 0$ esetén):

$$f_{Y_i}(y_i; \theta_i, \phi) = \\ \exp\left\{\frac{y_i(-\lambda_i/\alpha) - (-1)\log(\lambda_i/\alpha)}{1/\alpha} + (\alpha - 1)\log(y_i) - \log\Gamma(\alpha) + \alpha\log(\alpha)\right\}.$$

1.1.2. Általánosított lineáris modell a biztosításban

Tegyük fel, hogy van m megfigyelésünk, és jelölje az i-edik megfigyelés magyarázott változóját Y_i (kárdarabszám vagy átlagkár), további jellemzőit (nem, életkor stb.) pedig $X_{i,j}$. Ezeket a jellemzőket összegyűjthetjük egy X mátrixba, ahol a mátrix i-edik sora $X_i = \{X_{i,j} : j\}$ az i-edik biztosított jellemzőit gyűjti össze. $X_{i,j}$ lehet kategorikus és folytonos változó is, azonban én a dolgozatomban csak azzal az esettel foglalkozom, amikor minden magyarázó változó kategorikus. Ebben az esetben az X mátrix csak 0-1 elemeket tartalmaz: amennyiben az i-edik megfigyelés rendelkezik a j-edik tulajdonsággal, akkor $X_{i,j} = 1$, különben $X_{i,j} = 0$. Legyen például két magyarázó változónk, a szerződő neme (férfi, nő) és életkora (18-30, 31-50, 51+), ekkor 5 féle $X_{i,j}$ magyarázó változó van:

$X_{i,1}$: a szerződő férfi $X_{i,2}$: a szerződő nő $X_{i,3}$: a szerződő 18-30 éves
$X_{i,4}$: a szerződő 31-50 éves $X_{i,5}$: a szerződő 50 évnél idősebb.

Legyen 4 megfigyelésünk: egy 24 éves nő, egy 60 éves férfi, egy 50 éves nő, és egy 35 éves férfi. Ekkor:

$$X = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

Különböző magyarázó változók osztályainak metszetét *szegmensnek* nevezzük. A biztosítottakat a közös jellemzőik alapján ilyen szegmensekbe soroljuk. Vegyünk egy olyan példát, amikor 3 magyarázó változónk van: életkor (18-30, 31-50, 51+), nem (férfi, nő) és lakhely (A, B, C). Ekkor egy szegmenst alkotnak például a 18-30 év közötti, B-ben élő férfiak.

A modell az alábbi három fontos dolgot feltételezi.

1. *Szerződések függetlensége*: legyen n különböző biztosítási szerződésünk, és legyen az i-edik szerződés magyarázott változója Y_i. Ekkor $Y_1, \ldots, Y_n$ függetlenek.

2. *Időbeli függetlenség*: tegyük fel, hogy van n diszjunkt időintervallumunk, és legyen az i-edik intervallumbeli magyarázott változó Y_i. Ekkor $Y_1, \ldots, Y_n$ függetlenek.

3. *Homogenitás*: tegyük fel, hogy van két szerződésünk, amelyek ugyanabban a szegmensben helyezkednek el, és azonos ideig voltak kockázatban, vagy azonos számú kárt okoztak, továbbá a magyarázott változóik Y_1 és Y_2. Ekkor Y_1 és Y_2 azonos eloszlásúak.

A 3. tulajdonság alapján tehát egy adott szegmensben a biztosítottaknak egyforma a kárszükségletük.

Az exponenciális szórásmodellben gyakori választás az $a_i(\phi) = \phi/\omega_i$, ahol ω_i jelöli az i-edik megfigyelés súlyát, és így (1.3) alapján

$$\mathbb{D}^2(Y_i) = \frac{\phi \cdot V(\mu_i)}{\omega_i},$$

tehát az egyes megfigyelések súlyait is figyelembe tudjuk venni az általánosított lineáris modellben. Innentől a szakdolgozatomban csak ezt az $a_i(\phi) = \phi/\omega_i$ függvényt fogom használni.

Egy megfigyelés súlya lehet például a kárdarabszám, amennyiben a kárnagyságot szeretnénk modellezni, de a károk nem elérhetőek káreseményenként lebontva, csak szerződésenként összegezve. Például, ha csak annyit tudunk, hogy egy szerződésen 4 kár történt, amelyeknek az összege 100 000 Ft, akkor a modellben az átlagkárt vesszük figyelembe (ami jelen esetben 25 000 Ft), súlyként használva a kárdarabszámot. Ezzel tulajdonképpen azt modellezzük, mintha történt volna 4 darab 25 000 Ft értékű kár, ami így valóban összesen 100 000 Ft (persze a valóságban lehet hogy volt 2 darab 5000 Ft értékű és 2 db 45 000 Ft értékű, de sajnos az összesített adatokból ez már nem deríthető ki, és így ez a legjobb feltételezés, amivel becsülhetünk).

Fontos észrevétel, hogy az exponenciális szórásmodell reproduktív, azaz ha Y_1 és Y_2 független valószínűségi változók, amelyek ugyanahhoz az exponenciális szórásmodellhez tartoznak, és csak a súlyaik különböznek (ω_1 és ω_2), akkor a súlyozott átlaguk, $Y = (\omega_1 Y_1 + \omega_2 Y_2)/(\omega_1 + \omega_2)$ is ugyanahhoz az exponenciális szórásmodellhez tartozik, $\omega = \omega_1 + \omega_2$ súllyal. Ebből pedig az következik, hogy ha az általánosított lineáris modellben összevonjuk egy faktor két osztályát, feltételezve, hogy a megfigyeléseik azonos eloszlásúak, akkor az összevont csoport eloszlása is az exponenciális szórásmodell tagja lesz.

Az általánosított lineáris modellben az Y vektor minden eleme független, és az exponenciális szórásmodellhez tartozik. Az Y vektor és az X magyarázó változók mátrixa között a következő kapcsolat áll fenn:

$$\mu := \mathbb{E}(Y) = g^{-1}(\eta),$$

ahol $\eta = X \cdot \beta$ az ún. *lineáris prediktor*, $g(x)$ pedig az ún. *link függvény*, amely monoton és differenciálható, (így létezik $g^{-1}(x)$ az ún. *inverz link függvény*), és célunk a β paramétervektor becslése. Ezen egyenlőség és az (1.2) egyenlőség alapján

$$\mu = \mathbb{E}(Y) = b'(\theta) = g^{-1}(\eta). \tag{1.4}$$

Ha η helyére behelyettesítjük $X\beta$-t, azt kapjuk, hogy

$$b'(\theta) = g^{-1}(X\beta),$$

így tehát láthatjuk, hogy kapcsolat áll fenn a β és a θ paraméterek között.

Amennyiben egy magyarázó változó hatása már a becslés előtt ismert, akkor nem szeretnénk hozzá β paramétert becsülni, hanem inkább a rendelkezésre álló információt is szeretnénk hasznosítani a modellben. Az ilyen változókat *offszetnek* nevezzük, ξ-vel jelöljük, és egy ilyen változó hatása a következő módon illeszthető a modellbe:

$$\eta = X \cdot \beta + \xi.$$

Tehát ξ paraméterét nem becsüljük, hanem a priori 1-nek állítjuk be. Amennyiben a kárdarabszámot szeretnénk becsülni, és a link függvény logaritmikus (azaz $g(x) = \log x$), akkor a kockázatban töltött időt (d_i) a következőképpen vehetjük figyelembe:

$$\eta_i = X_i \cdot \beta + \log(d_i),$$

és ekkor

$$\mathbb{E}(Y_i) = e^{\eta} = d_i e^{X_i \cdot \beta}.$$

Ez tehát azt jelenti, hogy aki kétszer annyi időt töltött kockázatban, az várhatóan kétszer annyi kárt okoz, mint az, aki egységnyi időt volt kockázatban.

1.1.3. A modell struktúrája

Összefoglalva tehát az általánosított lineáris modell a következő (a szórásnégyzetnél feltételezve, hogy $a_i(\phi) = \phi/\omega_i$ alakú az exponenciális szórásmodellben):

$$\mathbb{E}(Y_i) = g^{-1}\left(\sum_j X_{i,j}\beta_j + \xi_i\right) = \mu_i$$

$$\mathbb{D}^2(Y_i) = \frac{\phi \cdot V(\mu_i)}{\omega_i}$$

ahol

Y_i a magyarázott változó vektor i-edik eleme,

$g(x)$ a link függvény,

$X_{i,j}$ a magyarázó változókból képzett mátrix i-edik sorának j-edik eleme,

β_j a paramétervektor j-edik eleme,

ξ_i az ismert hatások offszet vektorának i-edik eleme,

$V(x)$ a varianciafüggvény,

ϕ a szórásparaméter,

ω_i az i-edik megfigyelés súlya.

Célunk a β paramétervektor maximum likelihood becslése. A megfigyelések vektora $Y = \big(Y_1, \dots, Y_m\big)^T$, amelynek minden eleme független, és az exponenciális szórásmodellhez tartozik. Az exponenciális szórásmodellhez tartozó eloszlások sűrűségfüggvénye alapján a $\theta = \big(\theta_1, \dots, \theta_m\big)^T$ paramétervektorhoz tartozó likelihood függvény

$$L(\theta; \phi, y) = \prod_{i=1}^{m} f_{Y_i}(y_i; \theta_i, \phi) = \prod_{i=1}^{m} \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\},$$

és mivel $a_i(\phi) = \phi/\omega_i$, így

$$L(\theta; \phi, y) = \prod_{i=1}^{m} \exp\left\{ \frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \right\},$$

és ezek alapján a loglikelihood függény

$$\begin{aligned} \ell(\theta; \phi, y) &= \sum_{i=1}^{m} \frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \\ &= \frac{1}{\phi} \sum_{i=1}^{m} \omega_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{m} c(y_i, \phi). \end{aligned} \tag{1.5}$$

Mivel $c(y_i, \phi)$ nem függ θ-tól, így θ és β maximum likelihood becslésének kiszámításánál a deriváláskor ki fog esni.

Ahhoz, hogy a β paramétervektor maximum likelihood becslését megkaphassuk a θ-ra kapott loglikelihood függvény segítségével, felhasználjuk, hogy (1.4) alapján $\mu_i = b'(\theta_i)$ és

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} X_{i,j}\beta_j + \xi_i. \tag{1.6}$$

Vegyük észre, hogy nem számít, hogy ebben az (1.6) egyenletben szerepel-e a ξ_i offszet hatás, hiszen a β_j szerinti deriválás során úgyis kiesik. Így tehát a loglikelihood függvény β_j szerinti deriváltja a láncszabály alapján:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{m} \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{m} \omega_i(y_i - b'(\theta_i)) \frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{1}{\phi} \sum_{i=1}^{m} \omega_i(y_i - b'(\theta_i)) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \end{aligned} \tag{1.7}$$

Mivel $\mu_i = b'(\theta_i)$, így $\partial\mu_i/\partial\theta_i = b''(\theta_i)$. Kihasználva, hogy $V(\mu_i) = b''(\theta_i)$, az inverz függvény deriválási szabálya alapján $\partial\theta_i/\partial\mu_i = 1/V(\mu_i)$. Hasonlóan, mivel $\mu_i = g^{-1}(\eta_i)$, így $\partial\mu_i/\partial\eta_i = 1/g'(\mu_i)$. Továbbá mivel $\eta_i = \sum_j X_{i,j}\beta_j$, így $\partial\eta_i/\partial\beta_j = X_{i,j}$.

Amennyiben a kapott deriváltakat behelyettesítjük az (1.7) egyenletbe, azt kapjuk, hogy

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{m} \omega_i \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} X_{i,j}.$$

Ha ezeket a deriváltakat beszorozzuk ϕ-vel és egyenlővé tesszük nullával, megkapjuk a maximum likelihood egyenleteket:

$$\sum_{i=1}^{m} \omega_i \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} X_{i,j} = 0, \qquad j = 1, \ldots, p, \tag{1.8}$$

ahol p a becsülendő paraméterek számát jelöli. Fontos, hogy μ_i függ a β paramétervektortól, ugyanis (1.6) alapján

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\sum_{j=1}^{p} X_{i,j} \beta_j + \xi_i \right),$$

így ezt behelyettesítve a kapott maximum likelihood egyenletekbe, és megoldva β-ra, megkapjuk a β paramétervektor maximum likelihood becslését.

Fontos azonban megjegyezni, hogy a számítógépes programok iteratív eljárással határozzák meg a becsült β paramétereket, ugyanis nagy adatmennyiség esetén a megoldás pontos kiszámítása nagyon bonyolulttá válik. A leggyakrabban alkalmazott eljárás az ún. Newton–Raphson-módszer, ahol kiindulunk egy $\beta^{(0)}$ paramétervektorból (például $\beta^{(0)} = 0$), és az iteratív lépés a következő:

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1} \cdot s,$$

ahol H egy $p \times p$ méretű mátrix, amely a loglikelihood függvény második deriváltjait tartalmazza, s pedig egy p hosszúságú vektor, amely a loglikelihood függvény első deriváltjait tartalmazza. Amennyiben $\beta^{(n+1)}$ és $\beta^{(n)}$ eltérése kicsi, az iteráció megáll, és $\hat{\beta} = \beta^{(n+1)}$.

1.1.4. Gyakorlati megvalósítás

A modellben a β paramétervektort maximum likelihood becsléssel határozzuk meg. A szegmensek száma attól függ, hogy mennyi magyarázó változót, és azokon belül hány osztályt veszünk figyelembe (azaz, hogy milyen szinten aggregáljuk az adatokat). Amennyiben minden megfigyelést egy szegmensbe sorolunk (tehát nem vizsgáljuk a magyarázó változók hatását), akkor a becslésünk az egész portfolióra vonatkozó átlag lesz, ez az ún. *zérómodell* (angolul *null modell*). Ha viszont minden megfigyelésre külön becsülünk paramétereket (*teljes modell* vagy angolul *full modell*), akkor pontosan annyi β_i paramétert kapunk, ahány megfigyelésből áll a modell, és így könnyen lehetséges, hogy az egyenletrendszer túlhatározott lesz.

A kettő közti kompromisszumot adja az a módszer, amelyben az általunk kiválasztott magyarázó változók által meghatározott szegmensekre végezzük el a

becslést. Ezt úgy valósítjuk meg, hogy kijelölünk egy szegmenst, ez lesz az ún. *alaposztály* (legyen például a korábban említett, 18-30 év közötti B-beli férfiak szegmense), ennek a becsült paramétere legyen β_0 (ez az *alaposztály becslése*, angolul *intercept term*), és minden további jellemzőnek is becsülünk egy-egy β_i paramétert. Ekkor a magyarázó változókból alkotott X mátrix első oszlopa csupa 1-esekből áll (ez az oszlop felel meg az alaposztálynak), a további oszlopok pedig az alaposztálytól való eltérést mutatják. Például legyenek a megfigyeléseink a következőek: egy 24 éves B-beli nő, egy 60 éves A-beli férfi, egy 50 éves C-beli nő, és egy 35 éves C-beli férfi. Mivel az alaposztály a 18-30 év közötti B-beli férfiak szegmense, így most az X mátrix a következőképpen néz ki (a sorok rendre az említett megfigyelések):

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

ahol a második illetve a harmadik oszlop jelöli, ha a megfigyelés kora eltér az alaposztályétól (31-50 év közötti vagy 50 év feletti), a negyedik és ötödik oszlop jelöli, hogyha nem B-ben lakik (hanem A-ban vagy C-ben), az utolsó oszlop pedig azt jelöli, ha a megfigyelés nem férfi, hanem nő.

Az alaposztálybeli szerződések kárszükséglete β_0, és (logaritmikus link függvényt alkalmazva) minden további szegmens kárszükséglete $\beta_0 \cdot \beta_{i_1} \cdots \beta_{i_k}$, ahol az i-edik szegmensnek az alaposztálytól eltérő jellemzőinek becsült paraméterei a β_{i_j}-k.

Tehát, ha egy szerződés valamelyik jellemzője eltér az alaposztályétól, akkor annak a jellemzőnek a paraméterével még be kell szorozni β_0-t, és így kapjuk meg a szerződés kárszükségletét. A korábbi példán alkalmazva a következő lenne a paraméterezés:

Életkor	**Paraméter**
18-30	
31-50	β_1
50+	β_2

Lakhely	**Paraméter**
A	β_3
B	
C	β_4

Nem	**Paraméter**
Férfi	
Nő	β_5

Alaposztály becslése	β_0

Így tehát egy 18-30 év közötti B-beli férfi kárszükséglete β_0, de például egy 31-50 év közötti B-beli nő kárszükséglete $\beta_0 \cdot \beta_1 \cdot \beta_5$. Ezzel a módszerrel jóval kevesebb β_i paramétert kell becsülni, mint a teljes modell esetén, és így az egyenletrendszer egyértelműen meghatározott.

1.1.5. Illeszkedésvizsgálat

Az általánosított lineáris modell használatának egyik előnye, hogy hipotézisvizsgálattal tesztelni tudjuk, hogy az általunk készített modell mennyire

illeszkedik jól az adatokra. Az illeszkedés megfelelőségét a már korábban említett teljes modell segítségével vizsgálhatjuk, hiszen ez a modell tökéletesen illeszkedik az adatokra, így az általunk vizsgált modellt összehasonlíthatjuk a teljes modellel. Ezt az összehasonlítást segíti egy távolságfogalom, amely a vizsgált modell eltérését mutatja meg a teljes modelltől a likelihood-hányados próba segítségével.

Jelölje a becsült θ vektort $\tilde{\theta}$ a teljes modell esetén, $\hat{\theta}$ pedig a vizsgált modell esetén, továbbá jelölje $\hat{\mu}$ a vizsgált modellben becsült várható értékek vektorát. Ekkor, amennyiben a megfigyelések száma m, a likelihood-hányados próba az (1.5) egyenlőség alapján a következő:

$$\begin{aligned} 2 \cdot \Big(\ell(\tilde{\theta}; \phi, y) - \ell(\hat{\theta}; \phi, y)\Big) &= \frac{2}{\phi} \cdot \sum_{i=1}^{m} \omega_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - \Big(b(\tilde{\theta}_i) - b(\hat{\theta}_i)\Big)\right] \\ &= \frac{D(y, \hat{\mu})}{\phi}, \end{aligned}$$

ahol $D(y, \hat{\mu})$ jelöli a vizsgált modell távolságát a teljes modelltől.

Amennyiben a vizsgált modellben a becsült paraméterek száma p, akkor a likelihood-hányados próba eloszlása:

$$\frac{D(y, \hat{\mu})}{\phi} = 2 \cdot \Big(\ell(\tilde{\theta}; \phi, y) - \ell(\hat{\theta}; \phi, y)\Big) \sim \chi^2_{m-p}. \tag{1.9}$$

Gyakran azonban ϕ értéke nem ismert, és a modellezés során ezt is becsülni kell. Mivel (1.9) alapján

$$\mathbb{E}\left(\frac{D(y, \hat{\mu})}{\phi}\right) = m - p, \tag{1.10}$$

így egy gyakran használt becslés ϕ-re a következő:

$$\hat{\phi}_D = \frac{D(y, \hat{\mu})}{m - p}.$$

Ezt a távolságfogalmat alkalmazhatjuk egymásba ágyazott modellek összehasonlítására is. Két modellt akkor nevezünk egymásba ágyazottnak, ha az egyik modell magyarázó változóinak halmaza részhalmaza a másik modell magyarázó változói által alkotott halmaznak, vagy ha az egyik modellben egy adott magyarázó változó osztályainak halmaza részhalmaza a másik modellben ugyanazon magyarázó változó osztályai által alkotott halmaznak (vagy ha mindkettő teljesül). Az előbbi esetre példa, ha az A modellben a magyarázó változók a szerződő neme és az életkora, a B modellben pedig a szerződő neme, életkora és lakhelye, akkor az A modell a B modellbe van ágyazva. Az utóbbi esetre egy példa, ha az A modellben az életkor változónak 3 osztálya van (pl. 18-30, 31-50, és 50 év feletti), a B modellben pedig 2 osztálya van, amely részhalmaza az A-beli osztályoknak (pl. 18-50 és 50 év feletti), akkor a B modell az A modellbe van ágyazva. Azaz, például ha egy magyarázó változó osztályait összevonjuk, azzal egymásba ágyazott modelleket kapunk.

Tegyük fel, hogy az A modellben p_A a becsülendő paraméterek száma, a B modellben pedig p_B, és tegyük fel, hogy $p_B > p_A$, azaz az A modell a B modellbe van ágyazva. Ekkor, ha azt szeretnénk tesztelni, hogy az A modellt alkalmazhatjuk-e a B modell helyett egyszerűsítésképp, akkor használhatjuk a teljes modelltől való eltéréseik különbségeit tesztstatisztikának, azaz:

$$\frac{D(y,\hat{\mu}_A) - D(y,\hat{\mu}_B)}{\phi} = \frac{2[\ell_A(\tilde{\theta}_A;\phi,y) - \ell_B(\tilde{\theta}_B;\phi,y)]}{\phi} \sim \chi^2_{p_B-p_A}.$$

Így χ^2 próba segítségével tesztelhetjük azt a nullhipotézist, amely szerint a bővebb modellben az elhagyott $p_B - p_A$ darab paraméter mindegyike egyenlő nullával: amennyiben a tesztstatisztika értéke kisebb, mint a megfelelő kritikus érték, akkor azt mondhatjuk, hogy alkalmazhatjuk az A modellt a B modell helyett.

Egy másik lehetőség az illeszkedés vizsgálatára a Pearson-féle khi-négyzet próba, amelynek a tesztstatisztikája az általánosított lineáris modell esetén a következő:

$$\chi^2 = \sum_{i=1}^{m} \frac{(y_i - \hat{\mu}_i)^2}{\mathbb{D}^2(Y_i)} = \frac{1}{\phi} \sum_{i=1}^{m} \omega_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Hasonlóan az előző esethez, $\mathbb{E}(\chi^2) \approx m - p$, így ϕ becslése ebben az esetben:

$$\hat{\phi}_\chi = \frac{1}{m-p} \cdot \sum_{i=1}^{m} \omega_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

Amennyiben ϕ nem ismert, [1] alapján érdemesebb az utóbbi becslést alkalmazni, ugyanis $\hat{\phi}_D$ érzékenyebb a kerekítési hibákra.

1.2. Az aggregált károk modellje

Egy adott időszakban a biztosító által kifizetett kárkifizetés az abban az időszakban bekövetkezett és bejelentett károk összege. Az aggregált károk modellje az

$$S = \sum_{i=1}^{N} Y_i$$

véletlen tagszámú összeg, ahol N a károk száma, és $Y_1, \ldots, Y_N$ független azonos eloszlású kárnagyságok (tehát $Y_i \overset{d}{=} Y$).

A biztosító célja, hogy minden szerződésre megbecsülje ezeknek az aggregált károknak a várható értékét és szórását annak érdekében, hogy tisztában legyen egy-egy szerződő kockázatosságával.

1.2.1. Független eset

Tegyük fel, hogy a kárdarabszám és a kárnagyság független egymástól, azaz $Y_1, \ldots, Y_N$ nemcsak függetlenek és azonos eloszlásúak, hanem minden i-re Y_i független N-től. Ebben az esetben S eloszlása megkapható N és Y_i eloszlásaiból, és így ([1] alapján) az eloszlásfüggvénye $s \geq 0$ esetén

$$F_S(s) = \mathbb{P}(S \leq s) = \sum_{n=0}^{\infty} \mathbb{P}(S \leq s \mid N = n) \cdot \mathbb{P}(N = n),$$

továbbá felhasználva, hogy $Y_i \stackrel{d}{=} Y$, S generátorfüggvénye $(G_S(t))$, illetve momentumgeneráló függvénye $(M_S(t))$ a következő:

$$G_S(t) = G_N(G_Y(t)),$$

$$M_S(t) = M_N(\log(M_Y(t))). \tag{1.11}$$

S várható értéke és szórásnégyzete:

$$\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(Y), \tag{1.12}$$

$$\mathbb{D}^2(S) = \mathbb{E}^2(Y)\mathbb{D}^2(N) + \mathbb{E}(N)\mathbb{D}^2(Y), \tag{1.13}$$

azaz az összkár első két momentumát meghatározza a kárszám és a kárnagyság első két momentuma.

A biztosításban általában azt a feltételezést alkalmazzák, hogy a kárnagyság és a kárdarabszám független egymástól. Gyakran a kárdarabszámot $N \sim Poisson(\lambda)$ változóként, a kárnagyságot pedig $Y_i \sim \Gamma(\alpha, \beta)$ változóként modellezik. Ekkor S összetett Poisson-eloszlású, továbbá az $N = n$ feltétel mellett $S \sim \Gamma(n\alpha, \beta)$, és így az alábbiak teljesülnek:

$$F_S(s) = \begin{cases} \displaystyle\sum_{n=0}^{\infty} \int_0^s \frac{\beta^{\alpha n}}{y\Gamma(\alpha n)} y^{\alpha n} e^{-y\beta} \frac{\lambda^n e^{-\lambda}}{n!} dy & \text{ha } s > 0 \\ \mathbb{P}(N = 0) = e^{-\lambda} & \text{ha } s = 0 \end{cases}$$

$$\mathbb{E}(S) = \lambda \frac{\alpha}{\beta},$$

$$\mathbb{D}^2(S) = \lambda \left(\frac{\alpha + \alpha^2}{\beta^2} \right).$$

Továbbá, ebben az esetben N és Y momentumgeneráló függvénye

$$M_N(t) = \exp(\lambda(e^t - 1)) \quad \text{és} \quad M_Y(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha} \text{ ha } t < \beta,$$

amiből az (1.11) egyenlet alapján következik, hogy

$$M_S(t) = \exp[\lambda((1 - t/\beta)^{-\alpha} - 1)], \text{ ha } t < \beta.$$

Független esetben tehát a kárszükséglet várható értéke könnyen meghatározható: a kárnagyság és a kárdarabszám várható értékének szorzata. Ebben a modellben azonban nem tudjuk vizsgálni a kárszám és a kárnagyság között esetlegesen fennálló összefüggést. Az összefüggő esetben viszont van erre lehetőség, és ezáltal egy pontosabb becslés adható a biztosító várható kárkifizetésére.

1.2.2. Összefüggő eset

Legyenek továbbra is $Y_1, \ldots, Y_N$ függetlenek és azonos eloszlásúak adott N esetén, és tegyük fel, hogy minden i-re Y_i függ N-től, azaz a kárnagyság függ a kárdarabszámtól. Ebben az esetben S eloszlásának meghatározásához szükség van egy β_N paraméterre, ami megmutatja az összefüggést a kárszámok és a kárnagyságok között, és ekkor S várható értéke és szórásnégyzete már nem írható fel N és Y első két momentumából.

Független esetben ismert S eloszlása. Összefüggő esetben már sokkal bonyolultabb feladat felírni S eloszlásfüggvényét, így a kárszükséglet meghatározásához csak S várható értékét szeretnénk megbecsülni úgy, hogy az magában foglalja a kárszám és a kárnagyság közötti összefüggést. Ezt az általánosított lineáris modell segítségével tehetjük meg, felhasználva a kárdarabszám várható értékét, a kárnagyság feltételes várható értékét $N = n$ feltétel mellett, továbbá az említett β_N paramétert.

2. fejezet

Általánosított lineáris modell a független esetben

Vizsgáljuk most az aggregált károk modelljét a szerződések szintjén. Vegyük figyelembe az i-edik szerződő által okozott károk összegét:

$$S_i = \sum_{j=1}^{N_i} Y_{ij},$$

ahol N_i az i-edik szerződő által okozott károk száma, $Y_{i1}, \ldots, Y_{iN_i}$ pedig az általa okozott károk nagysága, amelyek független azonos eloszlásúak (tehát $Y_{ij} \stackrel{d}{=} Y_i$). Legyen továbbá az átlagos kárnagyság:

$$\overline{Y_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}.$$

Ekkor az aggregált károk modellje:

$$S_i = \sum_{j=1}^{N_i} Y_{ij} = N_i \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij} = N_i \overline{Y_i}.$$

Tehát az összkár a kárdarabszám és az átlagos kárnagyság (azaz átlagkár) szorzata.

Mivel a kárnagyságok függetlenek és azonos eloszlásúak, így az átlagkár várható értéke a következő:

$$\mathbb{E}(\overline{Y_i}) = \sum_{k=0}^{\infty} \mathbb{E}(\overline{Y_i} \mid N_i = k) \cdot \mathbb{P}(N_i = k) = \sum_{k=0}^{\infty} \mathbb{E}(Y_i \mid N_i = k) \cdot \mathbb{P}(N_i = k) = \mathbb{E}(Y_i)$$

Ez azt jelenti, hogy a tételes kárnagyságok és az átlagkár várható értéke megegyezik. Így az (1.12) egyenlőség alapján

$$\mathbb{E}(S_i) = \mathbb{E}(N_i)\mathbb{E}(\overline{Y_i}) = \mathbb{E}(N_i)\mathbb{E}(Y_i),$$

amiből pedig az következik, hogy az aggregált károk modellezésénél mindegy, hogy a tételes kárnagyságokat, vagy egy szerződés átlagos kárnagyságát alkalmazzuk. Szakdolgozatomban a modellezés során én az átlagkárt fogom használni.

Legyen az i-edik megfigyeléshez tartozó sor az X magyarázó változók mátrixában $x_i = \big(x_{i1}, \ \ldots, \ x_{ip}\big)$. Ekkor ha N_i és Y_i link függvénye g_{N_i} és g_{Y_i}, akkor

$$\nu_i := \mathbb{E}(N_i \mid x_i) = g_{N_i}^{-1}(x_i\alpha), \quad \text{és} \quad \mu_i := \mathbb{E}(Y_i \mid x_i) = g_{Y_i}^{-1}(x_i\beta),$$

ahol α és β p dimenziós oszlopvektorok, amelyek a becsült együtthatókat tartalmazzák. Fontos megjegyezni, hogy az i-edik megfigyelés magyarázó változói között szerepelhet egy ξ_i offszet hatás is, azonban annak az együtthatóját a priori 1-nek választjuk. Továbbá feltehető, hogy ν_i és μ_i ugyanazokból a magyarázó változókból állnak elő, ugyanis ha van olyan x_{ij}, amely csak az egyiknél fordul elő, akkor a neki megfelelő α_i-t vagy β_i-t választhatjuk a priori nullának a másiknál.

Mivel N_i és Y_i függetlenek tetszőleges szegmens esetén, így:

$$\mathbb{E}(S_i \mid x_i) = \mathbb{E}(N_i \mid x_i)\mathbb{E}(Y_i \mid x_i) = \nu_i\mu_i = g_{N_i}^{-1}(x_i\alpha) \cdot g_{Y_i}^{-1}(x_i\beta).$$

Azaz, az összkár várható értéke megkapható a kárdarabszám és a kárnagyság várható értékének szorzatából. Speciálisan, ha mindkét link függvény logaritmikus:

$$\mathbb{E}(S_i \mid x_i) = \nu_i\mu_i = e^{x_i\alpha + x_i\beta}. \tag{2.1}$$

Ekkor $\nu_i > 0$ és $\mu_i > 0$, tehát a kárdarabszám és a kárnagyság várható értéke pozitív, emiatt a biztosításban gyakran használják a logaritmikus link függvényt. A továbbiakban a szakdolgozatomban én is csak ezt a link függvényt fogom alkalmazni.

Amennyiben N_i és Y_i az exponenciális szórásmodellhez tartoznak, akkor (1.13) alapján

$$\mathbb{D}^2(S_i \mid x_i) = \nu_i\phi V_{Y_i}(\mu_i) + \psi V_{N_i}(\nu_i)\mu_i^2,$$

ahol V_{N_i} a kárdarabszámhoz, V_{Y_i} pedig a kárnagysághoz tartozó varianciafüggvény, és a hozzájuk tartozó szórásparaméterek ψ és ϕ.

A biztosításban gyakran feltételezik, hogy a kárdarabszám Poisson, a kárnagyság pedig Gamma eloszlású. Szakdolgozatomban a modellezés során én is ezekkel az eloszlásokkal fogok dolgozni. Amennyiben az i-edik megfigyelés kárdarabszámának eloszlása $N_i \sim Poisson(\nu_i)$, és az i-edik megfigyelés j-edik kárnagyságának eloszlása pedig $Y_{ij} \sim Gamma(1/\phi, 1/(\mu_i\phi))$, ahol $\nu_i > 0$ és $\mu_i > 0$, akkor S_i összetett Poisson-eloszlású, és ebben az esetben $V_{Y_i}(\mu_i) = \mu_i^2$, $V_{N_i}(\nu_i) = \nu_i$, $\psi = 1$, $\phi > 0$, így:

$$\mathbb{D}^2(S_i \mid x_i) = \phi\nu_i\mu_i^2 + \nu_i\mu_i^2 = \nu_i\mu_i^2(\phi + 1). \tag{2.2}$$

A továbbiakban csak ezt az esetet vizsgálom, amikor tehát N_i és Y_{ij} ilyen eloszlásúak.

A független esetben a kárdarabszámhoz és a kárnagysághoz tartozó becsülendő paraméterek α és β vektora külön-külön megbecsülhető az általánosított lineáris modell segítségével. Logaritmikus link függvénnyel a modellben a kárdarabszám és az átlagkár várható értéke

$$\nu_i = e^{x_i\alpha} \quad \text{és} \quad \mu_i = e^{x_i\beta},$$

ahol $\alpha = \big(\alpha_1, \ldots, \alpha_p\big)^T$ és $\beta = \big(\beta_1, \ldots, \beta_p\big)^T$ a becsülendő paramétervektorok, melyek azonosak minden i megfigyelésre. Tegyük fel, hogy az adataink m megfigyelést tartalmaznak. Ekkor (1.8) alapján α és β kiszámításához az alábbi maximum likelihood egyenleteket kell megoldani:

$$\sum_{i=1}^{m} \frac{(n_i - \nu_i)}{\nu_i} x_{ik}\nu_i = \sum_{i=1}^{m} x_{ik}(n_i - \nu_i) = 0, \qquad k = 1, \ldots, p,$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_i)}{\phi\mu_i^2} x_{ik}\mu_i = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{1}{\phi}\frac{x_{ik}}{\mu_i}(y_{ij} - \mu_i) = 0 \qquad k = 1, \ldots, p,$$

ahol n_i a megfigyelt kárdarabszámokat, y_{ij} pedig a megfigyelt kárnagyságokat jelöli.

Fontos észrevétel, hogy a külön-külön megfigyelt kárnagyságok (Y_{ij}) eloszlása nem egyezik meg az átlagkár $(\overline{Y_i})$ eloszlásával. Bár mindkettő Gamma eloszlású μ_i várható értékkel, a szórásnégyzeteik különböznek: $\mathbb{D}^2(Y_{ij}) = \mu_i^2\phi$, viszont az $N_i = n_i$ feltétel mellett $\mathbb{D}^2(\overline{Y_i}) = (\mu_i^2\phi)/n_i$.

Amennyiben az i-edik megfigyelés átlagkára $\overline{y}_i$, akkor az átlagkárral számolva a β paramétervektor kiszámításához az alábbi maximum likelihood egyenleteket kell megoldani:

$$\sum_{i=1}^{m} \frac{n_i}{\phi}\frac{x_{ik}}{\mu_i}(\overline{y}_i - \mu_i) = 0 \qquad k = 1, \ldots, p.$$

Némi átalakítással azonban a különálló kárnagyságokkal számolt β paramétervektorra vonatkozó maximum likelihood egyenletek is ugyanerre az alakra hozhatóak:

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{1}{\phi}\frac{x_{ik}}{\mu_i}(y_{ij} - \mu_i) &= \sum_{i=1}^{m} \frac{1}{\phi}\frac{x_{ik}}{\mu_i}\sum_{j=1}^{n_i}(y_{ij} - \mu_i) \\ &= \sum_{i=1}^{m} \frac{1}{\phi}\frac{x_{ik}}{\mu_i} n_i(\overline{y}_i - \mu_i) = 0 \qquad k = 1, \ldots, p.\end{aligned}$$

Ez tehát azt jelenti, hogy β becsléséhez mindegy, hogy a tételes kárnagyságokat vesszük figyelembe, vagy pedig az átlagkárt úgy, hogy a kárdarabszámot (n_i) használjuk súlyként. Szakdolgozatomban a modellezés során ez utóbbit fogom alkalmazni.

Fontos megjegyezni, hogy míg a kárdarabszámra vonatkozó modell a teljes adathalmazt figyelembe veszi, addig az átlagkár esetén csak azok a megfigyelések kerülnek bele a modellbe, amelyek esetén legalább 1 káresemény történt.

3. fejezet

Általánosított lineáris modell az összefüggő esetben

Vizsgáljuk ismét az aggregált károk modelljét a szerződések szintjén. Az összefüggő esetben azt feltételezzük, hogy az i-edik megfigyelés kárnagyságai (Y_{ij}) függnek a megfigyelés kárdarabszámától (N_i). Ez azt jelenti, hogy az $\mathbb{E}(\overline{Y_i} \mid N_i)$ feltételes várható érték (amely ekvivalens $\mathbb{E}(Y_{ij} \mid N_i)$-vel) az N_i kárdarabszám függvénye. A modellezés szempontjából ez azt jelenti, hogy az átlagkár becsléséhez a magyarázó változók közé vesszük a kárdarabszámot is.

A független esethez hasonlóan az összefüggő esetben is mindegy, hogy a kárnagyságokat (Y_{ij}) vagy az átlagkárt $(\overline{Y_i})$ használjuk a modellezés során. Azonban egy fontos különbség a két modell között, hogy míg független esetben az aggregált károk várható értéke felbontható a kárdarabszám és az átlagkár várható értékének szorzatára, addig az összefüggő esetben ez már nem teljesül. Ugyanis abban az esetben, ha N_i és $\overline{Y_i}$ összefüggőek, akkor:

$$\begin{aligned}\mathbb{E}(S_i \mid x_i) &= \mathbb{E}(N_i\overline{Y_i} \mid x_i) = \mathbb{E}\left(\mathbb{E}(N_i\overline{Y_i} \mid x_i, N_i) \mid x_i\right) \\ &= \mathbb{E}\left(N_i \cdot \mathbb{E}(\overline{Y_i} \mid x_i, N_i) \mid x_i\right) \\ &\neq \mathbb{E}(N_i \mid x_i)\mathbb{E}(\overline{Y_i} \mid x_i).\end{aligned}$$

Az a feltételezés, hogy az átlagkár függ a kárdarabszámtól, nincs hatással a kárdarabszám becslésére, tehát csakúgy, mint a független esetben, itt is

$$\nu_i = \mathbb{E}(N_i \mid x_i) = g_{N_i}^{-1}(x_i\alpha) = e^{x_i\alpha},$$

logaritmikus link függvényt alkalmazva. Azonban az átlagkár modellezése változik, hiszen ebben az esetben az összkár modellezéséhez szükség van az $\mathbb{E}(\overline{Y_i} \mid N_i, x_i)$ várható érték becslésére, amely a következő egyenlőség alapján történik:

$$g_{Y_i}\left(\mathbb{E}(\overline{Y_i} \mid N_i, x_i)\right) = x_i\beta + N_i\beta_N, \tag{3.1}$$

ahol $\beta = \big(\beta_1, \ldots, \beta_p\big)^T$ az eredeti p magyarázó változóhoz tartozó becsülendő paraméter, β_N pedig a kárdarabszámhoz tartozó paraméter. Ez a β_N együttható mutatja meg a kárdarabszám és az átlagkár közti összefüggést. Amennyiben β_N pozitív, az azt jelenti, hogy a nagyobb kárszámú megfigyelések átlagkára is nagyobb. Ha β_N negatív, akkor pont fordítva, a nagyobb kárszámú megfigyelésekhez kisebb átlagkár tartozik. Ha pedig $\beta_N = 0$, akkor a független esethez tartozó modellt kapjuk vissza.

Továbbra is logaritmikus link függvényt feltételezve, a (3.1) egyenlőség átrendezhető a következő alakra:

$$\mu_i^{(N)} := \mathbb{E}(\overline{Y_i} \mid N_i, x_i) = e^{x_i\beta + N_i\beta_N} \equiv \mu_i e^{\beta_N N_i}, \tag{3.2}$$

ahol tehát $\mu_i^{(N)}$ jelöli az összefüggő eset átlagkárának várható értékét (az (N) kitevővel utalva arra, hogy ebben az esetben már a kárdarabszám is a magyarázó változók között szerepel). Továbbá a független esethez hasonlóan μ_i jelöli azt az értéket, amely a kárdarabszámon kívül az összes többi magyarázó változó hatását tartalmazza. Ez a μ_i alakjában hasonlít a független modellben kapott átlagkár várható értékének becslésére, azonban itt az összefüggő modell β paramétervektora van behelyettesítve, ami nem egyezik meg a független esetben kapott β paramétervektorral, ugyanis a jelenlévő új β_N paraméter miatt a többi β_i paraméter becslése megváltozik az összefüggő modell esetén.

Így tehát a (3.2) egyenlőséget felhasználva, az aggregált károk várható értéke a következő:

$$\mathbb{E}(S_i \mid x_i) = \mathbb{E}\left(N_i \cdot \mathbb{E}(\overline{Y_i} \mid x_i, N_i) \mid x_i\right) = \mathbb{E}\left(N_i \mu_i e^{\beta_N N_i} \mid x_i\right) = \mu_i M'_{N_i}(\beta_N \mid x_i),$$

ahol M'_{N_i} jelöli N_i momentumgeneráló függvényének deriváltját a β_N helyen.

Amennyiben $N_i \sim Poisson(\nu_i)$, akkor $M_{N_i}(t) = \exp\{\nu_i(e^t - 1)\}$, és így az előzőek alapján:

$$\mathbb{E}(S_i \mid x_i) = \mu_i M'_{N_i}(\beta_N \mid x_i) = \nu_i \mu_i \exp\{\nu_i(e^{\beta_N} - 1) + \beta_N\}. \tag{3.3}$$

Ha ezt az eredményt összehasonlítjuk a (2.1) egyenlettel, láthatjuk, hogy az aggregált károk várható értékének becslésének alakja a független és az összefüggő esetben csak annyiban tér el, hogy az összefüggő esetben a képletben még szerepel egy $\exp\{\nu_i(e^{\beta_N} - 1) + \beta_N\}$ szorzó is, amire tekinthetünk úgy, hogy ez az összefüggésre vonatkozó korrekciós tag. Amennyiben $\beta_N = 0$, ez a korrekciós tag 1-gyel egyenlő, tehát visszakapjuk a független esetben kapott eredményt. Fontos azonban megjegyezni, hogy $\beta_N \neq 0$ esetén az összefüggő eset β paramétervektorának becslése nem egyezik meg a független eset β vektorának becslésével, tehát ekkor a független esetben kapott μ_i és a korrekciós tag szorzata nem egyenlő $\mu_i^{(N)}$-nel.

Összefüggő esetben (amennyiben $N_i \sim Poi(\nu_i)$, és $Y_{ij} \sim Gamma(1/\phi, 1/(\mu_i\phi))$, ahol $\nu_i > 0$ és $\mu_i > 0$) az aggregált károk szórásnégyzete a következő (a részletes

levezetés megtalálható az [1] irodalom 52-54. oldalán):

$$\begin{aligned}\mathbb{D}^2(S_i \mid x_i) = \nu_i\mu_i^2\Big[&\nu_i \exp\left\{\nu_i(e^{2\beta_N}-1)+4\beta_N\right\}\\ &+(\phi+1)\exp\left\{\nu_i(e^{2\beta_N}-1)+2\beta_N\right\}\\ &-\nu_i\exp\left\{\nu_i(e^{\beta_N}-1)+2\beta_N\right\}\Big].\end{aligned}$$

Amennyiben $\beta_N = 0$, azt kapjuk, hogy

$$\mathbb{D}^2(S_i \mid x_i) = \nu_i\mu_i^2\left[\nu_i\exp(0)+(\phi+1)\exp(0)-\nu_i\exp(0)\right] = \nu_i\mu_i^2(\phi+1),$$

amely megegyezik a független esetben kapott (2.2) egyenlőséggel.

Összességében tehát az összefüggő eset nagyon hasonló a független esethez, hiszen az aggregált károk várható értékét itt is a kárdarabszám és az átlagkár várható értékével (ν_i és $\mu_i^{(N)}$) tudjuk meghatározni, csak ebben az esetben a képletben még szerepel egy korrekciós szorzótényező, amely kifejezi az összefüggőséget. Csakúgy, mint a független esetben, itt is az általánosított lineáris modell segítségével határozzuk meg ν_i és $\mu_i^{(N)}$ várható értékét, az alábbi egyenletek alapján:

$$\nu_i = e^{x_i\alpha} \qquad \text{és} \qquad \mu_i^{(N)} = e^{x_i\beta+N_i\beta_N},$$

ahol tehát $\alpha = \big(\alpha_1,\ldots,\alpha_p\big)^T$, $\beta = \big(\beta_1,\ldots,\beta_p\big)^T$ és $\beta_N \in \mathbb{R}$ a becsülendő paraméterek. A független esetben α és β kiszámítható külön-külön az általánosított lineáris modell segítségével, azonban az összefüggő esetben egyben becsüljük meg az összes paramétert. Így a likelihood függvény felírásához szükségünk van az átlagkár és a kárdarabszám együttes sűrűségfüggvényére:

$$f_{\overline{Y},N}(\overline{y},n) = f_{\overline{Y}|N}(\overline{y} \mid n)\cdot f_N(n).$$

Így, amennyiben m megfigyelésünk van, a likelihood függvény a következő:

$$L(\alpha,\beta,\beta_N;y,n) = \prod_{i=1}^{m} f_{\overline{Y},N}(\overline{y}_i,n_i) = \prod_{i=1}^{m} f_{\overline{Y}|N}(\overline{y}_i \mid n_i)\cdot f_N(n_i),$$

ahol y_i és n_i jelöli a megfigyelt átlagkárokat és kárdarabszámokat, és így a loglikelihood függvény:

$$\ell(\alpha,\beta,\beta_N;y,n) = \sum_{i=1}^{m}\ell_N(\alpha;n_i) + \sum_{i=1}^{m}\ell_{\overline{Y}|N}(\beta,\beta_N;\overline{y}_i \mid n_i).$$

Láthatjuk, hogy a loglikelihood függvény felbomlik a kárdarabszám és az átlagkár loglikelihood függvényeinek összegére. Ezek alapján α-t az $\ell_N(\alpha;n_i)$ loglikelihood függvényből tudjuk megbecsülni, β-t és β_N-t pedig az $\ell_{\overline{Y}|N}(\beta,\beta_N;\overline{y}_i \mid n_i)$ függvényből. Amennyiben minden i esetén N_i Poisson-eloszlású ν_i várható értékkel, $\overline{Y}_i \mid N_i$ pedig Gamma eloszlású $\mu_i^{(N)}$ várható értékkel és $(\mu_i^2\phi)/N_i$ szórásnégyzettel,

akkor a megoldandó likelihood-egyenletek α-ra, β-ra és β_N-re rendre a következők (a részletes levezetés megtalálható az [1] irodalomban az 56-59. oldalon):

$$\sum_{i=1}^{m} x_{ik}(n_i - \nu_i) = 0 \qquad k = 1, \ldots, p,$$

$$\sum_{i=1}^{m} \frac{n_i}{\phi} \frac{x_{ik}}{\mu_i^{(N)}} (\overline{y}_i - \mu_i^{(N)}) = 0 \qquad k = 1, \ldots, p,$$

$$\sum_{i=1}^{m} \frac{n_i}{\phi} \frac{n_i}{\mu_i^{(N)}} (\overline{y}_i - \mu_i^{(N)}) = 0 \qquad k = 1, \ldots, p.$$

Fontos észrevétel, hogy összefüggő esetben az α-ra vonatkozó likelihood egyenletek megegyeznek a független esetben felírt egyenletekkel. Ezek alapján tehát α becslése megegyezik a függetlenséget és az összefüggést feltételező modell esetén. Ez azonban már nem mondható el a β paramétervektorra, hiszen itt az összefüggő esetben már a magyarázó változók között szerepel a kárdarabszám, ezáltal itt egy újabb paramétert is kell becsülni (β_N), ez pedig hatással van a többi β_i paraméterre is.

Összefoglalva tehát, az összefüggést feltételező modell egy viszonylag egyszerű kiterjesztése a független modellnek, hiszen ez utóbbiban annyi módosul, hogy az átlagkár modellezése során a magyarázó változók közé vesszük a kárdarabszámot. Ezáltal azonban a kárdarabszám becslése nem változik, és az összkár becslése továbbra is megkapható a kárdarabszám és az átlagkár várható értékének szorzatából, csak még be kell szorozni egy korrekciós tényezővel. Ráadásul, az átlagkár modellezésénél a magyarázó változóként használt kárdarabszám együtthatója megmutatja, hogy milyen kapcsolat áll fenn az átlagkár és a kárdarabszám között.

4. fejezet

Modellezés

Ebben a fejezetben egy konkrét példán keresztül vizsgálom a kárszámok és az átlagkár közti összefüggést, azaz alkalmazom az általánosított lineáris modellt a független és az összefüggő esetben is. Ezt az `R` program `glm` függvényével valósítom meg úgy, hogy az adatok 80%-át használom fel a modell illesztésre, és a maradék 20%-on vizsgálom az illeszkedést (ez az ún. *keresztkiértékeléses* módszer, angolul *cross-validation*).

4.1. Az adatok bemutatása

A modellezéshez a [6] irodalom *Car* nevű adathalmazát[1] használtam fel, amely egyéves, 2004 és 2005 közötti gépjármű biztosításokat tartalmaz. A megfigyelések száma 67 856, amelyek közül 4 624 esetben legalább 1 kár következett be (az átlagkár modellezéséhez csak az utóbbiakat használtam fel, míg a kárdarabszám modellezése során az összes megfigyelést felhasználtam).

A károk száma minden megfigyelés esetén 0-tól 4-ig terjedhet, eloszlásukat az alábbi táblázat mutatja be.

Kárszám	**Megfigyelések száma**	**Arány**
0	63 232	93,186%
1	4 333	6,386%
2	271	0,399%
3	18	0,027%
4	2	0,003%
Összesen:	*67 856*	*100%*

4.1. táblázat. A megfigyelt kárdarabszámok eloszlása

[1] Az adatok megtalálhatóak ezen a weblapon: `http://www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets`.

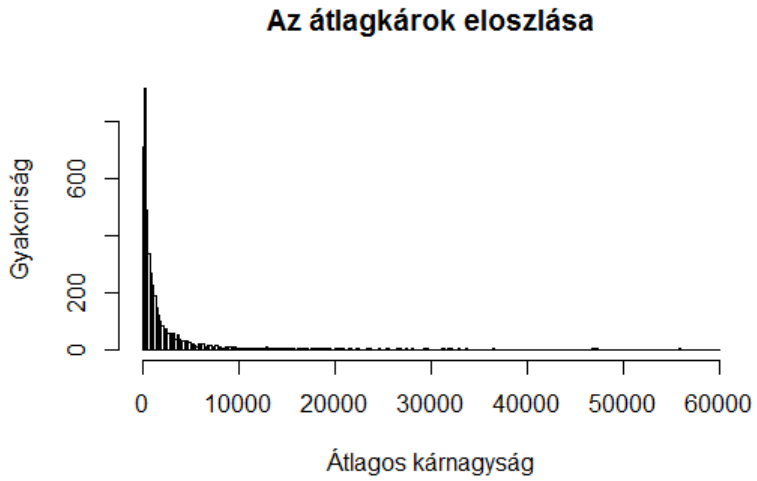
Láthatjuk, hogy azon szerződések száma, amelyek esetében több kár is történt, elég kevés. Egy-egy szerződés azonban különböző ideig volt kockázatban, így érdemes úgy is megvizsgálni a kárdarabszámok eloszlását, hogy nem a szerződések számát, hanem a kockázatban töltött időket adjuk össze. Az alábbi táblázat a kárszámonkénti összes kockázatban töltött időt tartalmazza (évben megadva).

Kárszám	**Összes kockázatban töltött idő**	**Arány**
0	28 974,299794	91,112%
1	2 619,780972	8,238%
2	192,232717	0,604%
3	12,736482	0,040%
4	1,768652	0,006%
Összesen:	*31 801*	*100%*

4.2. táblázat. A megfigyelt kárszámok eloszlása a kockázatban töltött idő szerint

Érdekesség, hogy ezen két táblázat alapján egy szerződő átlagos kockázatban töltött ideje körülbelül fél év. Továbbá láthatjuk, hogy az utóbbi esetben a kármentes esetek aránya csökkent, míg a pozitív károk aránya minden kárszámra nőtt. A gyakoriságok azonban itt és az előző esetben is arra utalnak, hogy a kárszámok Poisson-eloszlásúak, így a modellezés során ezt az eloszlást fogom feltételezni a kárdarabszámra. Fontos megjegyezni, hogy a biztosításban is legtöbbször Poisson-eloszlásúnak feltételezik a kárdarabszámot.

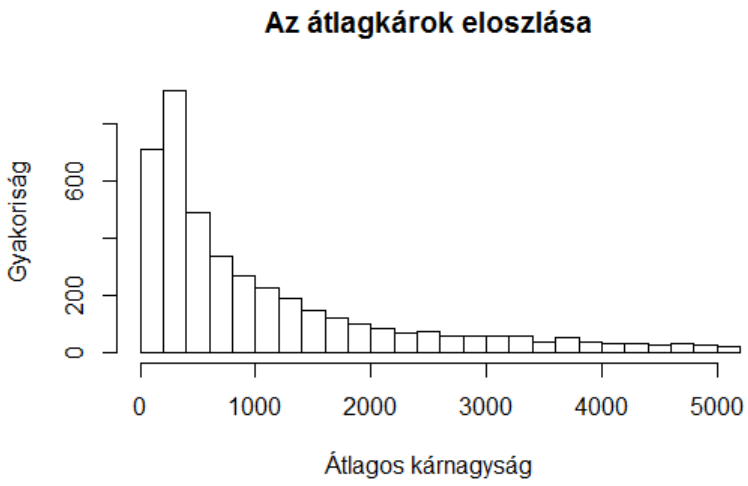
Az átlagkárok esetén a legkisebb érték 200, míg a legnagyobbé 55 922,13. Az átlagkárok eloszlását az alábbi ábra mutatja.



4.1. ábra. Az átlagkárok eloszlása a teljes átlagkár-terjedelemre

Láthatjuk, hogy a kis átlagkárok aránya elég magas, továbbá van néhány kiugró érték, azonban ezen az ábrán nehéz megvizsgálni, hogy pontosan milyen a kisebb

károk esetén az eloszlás. Ennek érdekében vizsgáljuk meg az alábbi ábrát, amely csak az 5 000 alatti átlagkárok eloszlását mutatja be.



4.2. ábra. Az átlagkárok eloszlása 5000 alatti átlagkárok esetén

Ezen ábra alapján feltehető, hogy az átlagkárok Gamma eloszlást követnek, így a modellezés során ezzel a feltételezéssel fogok élni. Fontos megjegyezni, hogy a biztosításban is legtöbbször Gamma eloszlásúnak feltételezik a kárdarabszámot.

Vizsgáljuk most meg, hogy melyek az egyes kárdarabszámhoz tartozó átlagkárok. Értelemszerűen 0 kár esetén a kárnagyság és így az átlagos kárnagyság is 0, a pozitív kárszámokhoz tartozó átlagkárokat (egészre kerekítve) pedig az alábbi táblázat mutatja.

Kárszám	**Átlagos kárnagyság ($)**
1	1 947
2	1 473
3	1 341
4	1 110

4.3. táblázat. Az átlagos kárnagyságok kárszámonként

Vegyük észre, hogy minél nagyobb a kárszám, annál kisebb az átlagos kárnagyság értéke. Ezek alapján feltételezhető, hogy az összefüggő esetben az átlagkár modellezése során (amikor a kárdarabszám is a magyarázó változók között szerepel) a kárszámhoz tartozó β_N együttható értéke negatív lesz.

Az `R` program `insuranceData` nevű csomagjában megtalálhatóak az eredeti adatok `dataCar` néven, én azonban átneveztem magyarra a változókat, így a modellemben az eredeti változók megfelelői a következőek (az eredeti adathalmazban szerepeltek olyan változók, amelyeket a modellezés során nem tudtam alkalmazni, így azok ebben a listában már nem szerepelnek):

`kockido`	**kockázatban töltött idő**; folytonos magyarázó változó; értéke legalább 0, legfeljebb 1 év
`karszam`	**károk száma**; diszkrét magyarázó és magyarázott változó; lehetséges értékei: 0,1,2,3,4
`karnagysag`	**összkár** (a szerződő által okozott károk összege); folytonos magyarázott változó; értéke legalább 0, legfeljebb 55 922,13 $
`gepj_tipus`	**gépjármű típusa**; kategorikus magyarázó változó, 13 féle gépjármű kategóriával
`gepj_erteke`	**gépjármű értéke**; folytonos magyarázó változó; értéke legalább 0, legfeljebb 34,56 (10 000 $-ban mérve)
`gepj_kor`	**gépjármű kora**; kategorikus magyarázó változó, 4 kategóriával
`nem`	**szerződő neme**; kategorikus magyarázó változó; nő vagy férfi
`kor`	**szerződő életkora**; kategorikus magyarázó változó, 6 kategóriával
`lakhely`	**szerződő lakhelye**; kategorikus magyarázó változó, 6 kategóriával

Mivel szerződésenként csak az összkár elérhető, így a modellezéshez az összkárból és a károk számából kiszámítottam egy átlagkár változót (`atlagkar`). Így magyarázott változó lesz a kárdarabszám és az átlagkár, a többi változó pedig magyarázó változó lesz. A modellezés során csak kategorikus változókat szeretnék alkalmazni, így a `gepj_erteke` változóból készítettem egy kategorikus változót (`gepj_erteke_kat`) úgy, hogy 4 különböző kategóriába soroltam az eredeti értékeket. A kockázatban töltött időt (`kockido`) viszont meghagytam folytonos változónak, mert ezt csak offszetként fogom alkalmazni a kárdarabszám modellezésénél (az 1.1.2. fejezetben leírtak alapján).

A modellezés során az adatok 80%-ára illesztettem a modellt, és a maradék 20%-on vizsgáltam az illeszkedést. Ehhez létrehoztam egy `datatype` nevű változót, amelynek értéke véletlen mintavételezés alapján az adatok 80%-ában „`training`”, a maradék 20%-ban pedig „`test`”.

4.2. Modellek a független esetben

Ebben a fejezetben bemutatom, hogy hogyan modelleztem a kárdarabszám és az átlagkár várható értékét abban az esetben, ha ezeket függetlennek feltételezzük. Megmutatom, hogy milyen `R` kódokkal valósítottam meg az egyes modelleket, és leírom, hogy milyen eredményeket kaptam az egyes modellek esetén.

4.2.1. Kárszám modell

A kárdarabszám modellezése során feltételeztem, hogy a kárszám Poisson-eloszlású, továbbá logaritmikus link függvényt alkalmaztam az általánosított lineáris modellben, és offszetként használtam a kockázatban töltött idő logaritmusát. Először belevettem a modellbe minden magyarázó változót, így az első kárdarabszám modellem `R` kódja a következő volt:

```
glm(karszam ~ gepj_erteke_kat + gepj_tipus
              + gepj_kor + nem + lakhely + kor,
    family = poisson(link = log),
    data   = adatok,
    subset = (datatype == "training" & kockido > 0),
    offset = log(kockido))
```

Ebben az esetben a `nem` változó egyáltalán nem volt szignifikáns, így ezt kivettem a modellből. Továbbá a többi változónak is voltak olyan osztályaik, amelyek nem voltak szignifikánsak, így összevontam bizonyos osztályokat úgy, hogy végül minden osztály szignifikáns legyen. Így tehát a végső modellem annyiban tért el az elsőtől, hogy nem szerepelt benne a `nem` változó, a többi magyarázó változóban pedig összevonásra kerültek bizonyos osztályok.

A végső modellben 11 paramétert becsültem, és így az i-edik szerződőhöz tartozó kárdarabszám várható értékének becslése a következő lett (a 2. fejezet jelöléseit alkalmazva):

$$\begin{aligned}\hat{\nu}_i = \texttt{kockido} \cdot \exp\big\{&\alpha_0 + \alpha_1 \cdot \texttt{gepj\_erteke\_kat}_1 + \alpha_2 \cdot \texttt{gepj\_erteke\_kat}_2 + \\ &\alpha_3 \cdot \texttt{gepj\_tipus}_A + \alpha_4 \cdot \texttt{gepj\_tipus}_B + \alpha_5 \cdot \texttt{gepj\_tipus}_C + \\ &\alpha_6 \cdot \texttt{gepj\_kor}_1 + \alpha_7 \cdot \texttt{lakhely}_A + \alpha_8 \cdot \texttt{kor}_1 + \alpha_9 \cdot \texttt{kor}_2 + \\ &\alpha_{10} \cdot \texttt{kor}_3\big\}.\end{aligned}$$

Itt tehát például a $\texttt{lakhely}_A$ változó értéke 1, ha az i-edik szerződő A-ban lakik, különben 0. A becsült paramétereket az alábbi táblázat tartalmazza:

α_0	-1,83338
α_1	-0,11042
α_2	0,14964
α_3	0,42091
α_4	0,73473
α_5	-0,20747
α_6	0,07688
α_7	-0,11792
α_8	0,23215
α_9	-0,26067
α_{10}	-0,19187

4.4. táblázat. A kárdarabszám modell esetén becsült paraméterek

Ezek alapján az alaposztálybeli szerződők kárdarabszámának várható értéke $e^{-1,83338} \approx 0,16$. Ha azonban tekintünk egy olyan szerződőt, akinek minden tulajdonsága megegyezik az alaposztálybeliekével, kivéve, hogy a gépjárműve egy másik, a legolcsóbb kategóriába tartozik ($\texttt{gepj\_erteke\_kat}_1$), akkor az ő kárdarabszámának a várható értéke $e^{-1,83338-0,11042} \approx 0,14$.

Érdekesség, hogy a legfiatalabb szerződők ($\texttt{kor}_1$) együtthatója pozitív, míg a két legidősebb csoport ($\texttt{kor}_2$ és $\texttt{kor}_3$) együtthatója negatív ($\texttt{kor}_0$ jelöli az alaposztályt, amelynek tagjai korban az 1-es és a 2-es csoport között helyezkednek el). Ez tehát azt jelenti, hogy a legfiatalabbak okozzák a legtöbb kárt, és az idősebbek pedig kevesebbet okoznak, mint a középkorúak. Nem mondható el viszont, hogy minél idősebb valaki, annál kevesebb kárt okoz, hiszen a legidősebb csoport ($\texttt{kor}_3$) együtthatója már nagyobb, mint az eggyel fiatalabb csoporté ($\texttt{kor}_2$), tehát a legidősebb korosztály már kicsivel több kárt okoz, mint az eggyel fiatalabb csoport szerződői.

A végső modell eltérése a teljes modelltől 20 185, míg a null modell eltérése a teljes modelltől 20 326. Mivel a teljes modell tökéletesen illeszkedik az adatokra, így az általam tesztelt modell jobbnak bizonyul, mint a null modell, hiszen kevésbé tér el a teljes modelltől.

A program ennél a modellnél $\phi = 1$ szórásparaméterrel számolt, ugyanis a Poisson-eloszlás esetén $a_i(\phi) = 1 = \phi/\omega_i$, és mivel itt minden megfigyelést azonos súllyal veszünk figyelembe, így $\omega_i = 1$ minden i esetén. Fontos azonban megjegyezni, hogy (1.10) alapján $D(y, \hat{\nu})/(m - p) \approx \phi$, ahol tehát $D(y, \hat{\nu})$ a modell távolsága a teljes modelltől, m a megfigyelések száma, p pedig a becsült paraméterek száma. Jelen esetben azon megfigyelések kerültek a modellbe, amelyeknek az adattípusa „training", a kockázatban töltött idejük pedig nagyobb, mint 0, így ennél a modellnél $m = 54\ 121$, tehát ebben az esetben

$$\frac{D(y, \hat{\nu}_i)}{m - p} = \frac{20\ 185}{54\ 121 - 11} = 0,373036407 \neq 1 = \phi.$$

Fontos azonban, hogy ebből még nem következik, hogy az adatok nem Poisson-eloszlásúak, ugyanis ez csak egy durva becslés ϕ-re, hiszen $D(y, \hat{\nu})/(m - p)$ eloszlását nem ismerjük. Ugyanakkor lehetséges, hogy a kárdarabszámra vonatkozó Poisson-eloszlás feltételezése nem a legmegfelelőbb. Mivel az adatokban nagy mennyiségű megfigyelés esetén 0 a kárszám, így elképzelhető, hogy egy olyan Poisson-eloszlással, amely keverve van az azonosan nulla eloszlással (például valamely $0 < \alpha < 1$ esetén $Y \sim \alpha \cdot Poisson(\nu) + (1-\alpha) \cdot 0$) megfelelőbb illeszkedést kapnánk.

4.2.2. Átlagkár modell

Az átlagkárok modellezése során Gamma eloszlást feltételeztem, és itt is logaritmikus link függvényt alkalmaztam, továbbá súlyként használtam a megfigyelések kárdarabszámát. Először itt is belevettem a modellbe minden magyarázó változót, így az első átlagkár modellem `R` kódja a következő volt:

```
glm(atlagkar ~ gepj_erteke_kat + gepj_tipus + gepj_kor
              + nem + lakhely + kor,
    family  = Gamma(link = log),
    data    = adatok,
    subset  = (karszam > 0 & datatype == "training"),
    weights = karszam)
```

Ebben az esetben sem volt minden változó szignifikáns, továbbá itt is összevontam a magyarázó változók egyes osztályait, hogy minden osztály külön-külön is szignifikáns legyen. Így a végső modellembe csak a következő változók kerültek be: `gepj_tipus`, `nem`, `lakhely` és `kor`.

A végső modellben 6 darab β_i paramétert becsültem, így az i-edik szerződő átlagkárának várható értéke a következő lett:

$$\hat{\mu}_i = \exp\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1\}.$$

A becsült paramétereket pedig az alábbi táblázat tartalmazza:

β_0	7,34311
β_1	0,15692
β_2	-0,68300
β_3	0,18000
β_4	0,39995
β_5	0,29137

4.5. táblázat. Az átlagkár modell esetén becsült paraméterek

Ezek alapján az alaposztálybeli szerződők átlagkárának várható értéke $e^{7,34311} \approx 1\,546$ \$. Az alaposztálybeli szerződők mind nők, azonban ha tekintünk egy szerződőt, aki férfi, de minden egyéb tulajdonsága megegyezik az alaposztálybeliekével, akkor az ő átlagkára várhatóan $e^{7,34311+0,18} \approx 1\,850$ \$ lesz. Tehát ezen modell alapján a férfiak várhatóan nagyobb károkat okoznak, mint a nők. Érdekesség továbbá, hogy a legfiatalabb korcsoport ($\texttt{kor}_1$) együtthatója pozitív, tehát a legfiatalabb korosztály várhatóan $e^{0,29137} \approx 1,34$-szer nagyobb károkat okoz, mint az idősebbek.

A végső modell eltérése a teljes modelltől 5 860, míg a null modell eltérése a teljes modelltől 5 999, így az általam tesztelt modell jobbnak bizonyul, mint a null modell.

4.2.3. Aggregált károk modellje

Egy szerződő összkárának várható értéke megkapható a kárdarabszám és az átlagkár várható értékének szorzatából. Így tehát az i-edik szerződőre vonatkozó

aggregált károk modellje a következő:

$$\mathbb{E}(S_i \mid x_i) = \mathbb{E}(N_i \mid x_i) \cdot \mathbb{E}(\overline{Y_i} \mid x_i) = \hat{\nu}_i \cdot \hat{\mu}_i$$

Ahhoz, hogy később össze tudjam hasonlítani a független eset és az összefüggő eset aggregált kármodelljeit, kiszámoltam, hogy mennyi a becsült összkárok és a valódi összkárok átlagos négyzetes, illetve átlagos abszolút eltérése. Ehhez csak azon adatokat vettem figyelembe, amelyeket csak tesztelésre szántam (ezek száma 13 735), és nem vettem figyelembe azon megfigyeléseket, amelyekből a modellt építettem fel. A valódi összkárokat az eredeti `karnagysag` változó tartalmazza, jelölje ezt minden szerződésre S_i. Így a következő eredményeket kaptam:

$$\sum_{i=1}^{13\ 735} \frac{|S_i - \hat{\nu}_i\hat{\mu}_i|}{13\ 735} = 250,8657 \qquad \sum_{i=1}^{13\ 735} \frac{(S_i - \hat{\nu}_i\hat{\mu}_i)^2}{13\ 735} = 1\ 083\ 092.$$

Hasonlóan ki fogom számolni ezeket az eltéréseket az összefüggő esetben is, ezáltal össze lehet majd hasonlítani, hogy melyik modell becslései térnek el kevésbé az eredeti összkároktól, azaz, hogy a teszt adatbázison melyik modell jelzi jobban előre az összkárt.

4.3. Modellek az összefüggő esetben

Ebben a fejezetben végig feltételezem, hogy a kárdarabszám és az átlagkár összefüggésben áll egymással, és bemutatom, hogy milyen eredményeket kaptam a különböző modellezések során. Ahogy a független esetben, úgy itt is Poisson-eloszlásúnak feltételezem a kárdarabszámot, és Gamma eloszlásúnak az átlagkárt.

4.3.1. Kárszám modell

Ahogy azt korábban láthattuk, a kárdarabszám modellezésénél nem számít, hogy függetlennek, vagy összefüggőnek feltételezzük a kárszámot és az átlagkárt. Így tehát az összefüggő esetben a kárdarabszámra vonatkozó modell megegyezik a független eset modelljével (4.2.1. fejezet).

4.3.2. Átlagkár modell

Az átlagos kárnagyság modellezése során több modellt is kipróbáltam. Kezdetben a független eset végső átlagkár modelljéből indultam ki, és a magyarázó változók közé vettem a kárdarabszámot, mint folytonos változót (ez nem feltétlenül a legjobb modell, azonban így könnyebb összehasonlítani a független és az összefüggő modellt). Lehetséges azonban, hogy egy jobb modellt kapunk, ha a kárdarabszám helyett az következők valamelyikét választjuk magyarázó változónak:

- log(`karszam`): Ha a kárszám helyett a kárszám logaritmusát vesszük figyelembe, akkor az átlagos kárnagyság várható értékének képletében nem $e^{\beta_N \cdot karszam}$, hanem $karszam^{\beta_N}$ fog szerepelni szorzóként. Ezáltal egy adott kárszámnak nem exponenciális, hanem hatvány hatása lesz az átlagkárra, amely lehet, hogy jobban leírja a valóságot.

- `kargyakorisag` (=`karszam`/`kockido`): Amennyiben minden megfigyelés esetén elosztjuk a kárdarabszámot a kockázatban töltött idővel, megkapjuk az adott megfigyelés kárgyakoriságát. Ha a kárszám helyett a kárgyakoriságot tesszük a modellbe magyarázó változóként, akkor a modellezés során már azt is figyelembe vesszük, hogy a szerződő mennyi idő alatt okozta az adott kárszámot, és ezáltal lehetséges, hogy egy jobb modellt kapunk.

- log(`kargyakorisag`): Ez az eset kombinálja az előző kettőt, ezáltal hatvány hatásként vesszük figyelembe, hogy egységnyi idő alatt ki mennyi kárt okozott.

- `karszam`, mint kategorikus változó: Ha folytonos változó helyett kategorikusként vizsgáljuk a kárdarabszámot, akkor ahelyett, hogy egy közös β_N paramétert becsülnénk a kárszámnak, külön-külön becsülünk paramétert a 0,1,...,4 kárdarabszámra. Ebben az esetben, ha például az 1 kárt és a 3-4 kárt okozó szerződők kisebb károkat okoznak, mint a 2 kárdarabszámú szerződők, akkor egy jobb modellt kaphatunk, hiszen egy ilyen esetre nem lehet jól modellt illeszteni, ha minden kárszámra egyetlen közös paraméter szerepel a modellben.

Mindegyik esetre felírtam egy-egy általánosított lineáris modellt az `R` programban. A továbbiakban bemutatom ezeket a modelleket, és megvizsgálom, hogy melyik tér el legkevésbé a teljes modelltől. A későbbiekben pedig azt az átlagkár modellt fogom választani az aggregált károk modelljéhez, amelyik a legközelebb van a teljes modellhez.

1. Modell

Elsőként tehát azt az esetet vizsgáltam, amikor egyszerűen a kárdarabszám, mint folytonos változó szerepel az átlagkár magyarázó változói között. Így tehát csak ki kellett egészítenem a független eset átlagkár modelljét a `karszam` változóval, amelyet az alábbi `R` kóddal valósítottam meg:

```
glm(atlagkar ~ gepj_tipus + nem + lakhely
               + kor + karszam,
    family   = Gamma(link = log),
    data     = adatok,
    subset   = (karszam > 0 & datatype == "training"),
    weights  = karszam)
```

Ezáltal egy szerződő átlagkárának várható értékének becslése a következő:

$$\hat{\mu}_i = \exp\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1 + \beta_6 \cdot \texttt{karszam}\}.$$

A becsült paramétereket pedig a következő táblázat tartalmazza:

β_0	7,57651
β_1	0,15262
β_2	-0,66054
β_3	0,17621
β_4	0,40285
β_5	0,28663
β_6	-0,20458

4.6. táblázat. Az 1. modell esetén becsült paraméterek

Láthatjuk, hogy a kárdarabszámhoz tartozó paraméter negatív, tehát minél több kárt okoz egy szerződő, várhatóan annál kisebb lesz az átlagkára (ez megfelel a 4.3 táblázat alapján megállapított feltételezésnek). Továbbá fontos, hogy a `karszam` változó p-értéke a program szerint 0,00412, tehát ez a változó szignifikáns a modellben 1%-os szignifikanciaszinten.

Független esetben az átlagkár modell eltérése a full modelltől 5 860 volt, itt az összefüggő esetben pedig 5 836. Ezek alapján az összefüggő eset modellje bizonyul jobbnak, amiből pedig arra lehet következtetni, hogy érdemes összefüggést feltételezni a kárdarabszám és az átlagos kárnagyság között.

2. Modell

Ebben az esetben a kárdarabszám helyett annak logaritmusát tettem az átlagkár magyarázó változói közé, így az `R` kódom csak annyiban tér el az előző esettől, hogy `karszam` helyett log(`karszam`) került a modellbe. Ezáltal az átlagkár várható értéke a következő:

$$\hat{\mu}_i = \texttt{karszam}^{\beta_6} \cdot \exp\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1\}.$$

Ebben az esetben a log(`karszam`) változó p-értéke a program szerint 0,003284, amiből az következik, hogy ha a kárszám logaritmusát tekintjük, az még szignifikánsabb, mintha egyszerűen csak a kárszámot tennénk a modellbe. Ugyanakkor 2. modell eltérése a full modelltől 5 835, amely bár egy kicsivel jobb, mint az 1. modell esetén, de nem érdemes részletesen vizsgálni a kapott együtthatókat. Összességében viszont elmondható, hogy a 2. modell jobb, mint az 1. modell.

3. Modell

Ebben az esetben a kárdarabszámot, mint kategorikus változót tettem az átlagkár magyarázó változói közé (ezáltal tehát a modell nem egy közös β_6 paramétert becsül a kárdarabszámra, hanem külön-külön β_i paramétert minden egyes kárszámra). A modellt a következő `R` kóddal valósítottam meg:

```
glm(atlagkar ~ gepj_tipus + nem + lakhely
               + kor + as.factor(karszam),
    family  = Gamma(link = log),
    data    = adatok,
    subset  = (karszam > 0 & datatype == "training"),
    weights = karszam)
```

Ebben az esetben azonban a 3 és a 4 kár kategóriája nem volt szignifikáns (amely nem meglepő, hiszen a 3 és 4 kárdarabszámmal rendelkező megfigyelések száma elég kevés), így készítettem egy változót, amelyben összevontam ezeket a kategóriákat a 2 kár kategóriájával. Így a végső modellben egy olyan kategorikus kárdarabszám változó szerepelt, amelynek 2 kategóriája van: a szerződő 1 kárt okozott, vagy 1-nél több kárt (az alaposztályba azok kerültek, akik 1 kárt okoztak). Ebben az esetben azonban a modell eltérése a full modelltől 5 836, amely ugyanannyi, mint az 1. modell esetén, azaz amikor a kárdarabszám folytonos változóként szerepel.

Összességében tehát elmondható, hogy nem javít a modellezésen, ha a kárdarabszámot kategorikus változóként alkalmazzuk.

4. Modell

Ebben az esetben a `karszam` helyett a `kargyakorisag` változó hatását vizsgáltam. Így tehát az i-edik szerződés átlagkárának várható értéke a következő:

$$\hat{\mu}_i = \exp\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1 + \beta_6 \cdot \texttt{kargyakorisag}\}.$$

Ebben a modellben a `kargyakorisag` változó p-értéke a progam szerint $1,82 \cdot 10^{-13}$, ami azt jelenti, hogy így egy sokkal szignifikánsabb változót kapunk, mint az 1. és 2. modell esetén a `karszam`, illetve a log(`karszam`) változó alkalmazásával. Továbbá a modell eltérése a teljes modelltől 5 780, amely már lényegesen kisebb, mint a 2. és az 1. modell esetén kapott eltérések. Ezek alapján érdemes megvizsgálni a kapott együtthatókat is, összehasonlítva az eredeti (1.) modell értékeivel. A következő táblázat tartalmazza a 1. modell és a 4. modell együtthatóit.

Paraméter	**1. modell**	**4. modell**
β_0	7,57651	7,26733
β_1	0,15262	0,17260
β_2	-0,66054	-0,65488
β_3	0,17621	0,18880
β_4	0,40285	0,37889
β_5	0,28663	0,26044
β_6	-0,20458	0,01975

4.7. táblázat. Az 1. és a 4. modell esetén becsült paraméterek

Láthatjuk, hogy az alaposztály becslése (β_0) a 4. modell esetén egy kicsit csökkent, továbbá a két modellben megegyező változókhoz tartozó paraméterek (β_1-β_5) is változtak némileg. A legfontosabb eltérés azonban, hogy míg az 1. modell esetén a kárszámhoz tartozó β_6 paraméter negatív, addig a 4. modellben a kárgyakorisághoz tartozó β_6 paraméter pozitív. Ebből az következik, hogy minél nagyobb egy szerződő kárgyakorisága, annál nagyobb károkat okoz.

Mivel az [1] és [2] irodalmak csak a kárszámot használták magyarázó változóként, és ezen irodalmakból indultam ki a modellezésem során, így az adatok bemutatásakor csak azt vizsgáltam, hogy az átlagkárok hogyan oszlanak el a kárszámok szerint. Mivel azonban a 4. modellem szerint sokkal jobb illeszkedést mutat, ha a kárgyakoriságot használjuk magyarázó változóként, ráadásul ellentétes előjelű hatása is van, érdemes megvizsgálni, hogy hogyan oszlanak el az átlagkárok a kárgyakoriságok szerint.

A kárgyakoriság azt mutatja meg, hogy egy szerződő egy év alatt hány kárt okoz, így az értéke nulla minden olyan megfigyelés esetén, ahol nem történt kár. Azon megfigyelésekre, ahol pedig legalább egy kár történt, ott a kárgyakoriság azt mutatja meg, hogy egy szerződő egy év alatt várhatóan hány kárt okoz (ez az érték pedig legalább 1, hiszen a nevezőben a kockázatban töltött idő legfeljebb 1). Tehát azzal, hogy a kárgyakoriságot vizsgáljuk, tulajdonképpen mindenki kárszámát standardizáljuk azonos kockázatban töltött időszakra.

Mivel a `kargyakorisag` nem kategorikus változó, így beosztottam az értékeit bizonyos intervallumokba, és az alábbi táblázat azt mutatja be, hogy mennyi a szerződők átlagkára ezekben az intervallumokban.

Kárgyakoriság (x)	**Átlagos kárnagyság (\$-ban)**
$x < 1$	0
$1 \leq x < 1,5$	1 490
$1,5 \leq x < 2,5$	1 881
$2,5 \leq x < 3,5$	1 888
$3,5 \leq x$	2 744

4.8. táblázat. Az átlagos kárnagyságok az egyes kárgyakoriság intervallumokban

Láthatjuk, hogy valóban igaz, amit a modell is becsült, azaz, hogy a nagyobb kárgyakoriságú szerződőknek nagyobb az átlagos kárnagysága.

Fontos továbbá, hogy míg 2-nél több kárszámmal rendelkező megfigyelés alig található az adatok között (a 3 és 4 kárdarabszámmal rendelkező szerződések aránya összesen 0,029%, ld. 4.3 táblázat), addig a 2-nél nagyobb kárgyakorisággal rendelkező megfigyelések száma igen jelentős, amelyet megfigyelhetünk a következő táblázat alapján (ez annak köszönhető, hogy az átlagos kockázatban töltött idő 0,5 év, és így a kárgyakoriságok átlagosan dupla akkorák, mint a megfigyelt kárdarabszámok).

Kárgyakoriság (x)	**Megfigyelések száma**	**Arány**
$x < 1$	63 232	93,186%
$1 \leq x < 1,5$	1 710	2,520%
$1,5 \leq x < 2,5$	983	1,449%
$2,5 \leq x < 3,5$	975	1,437%
$3,5 \leq x$	956	1,409%
Összesen:	*67 856*	*100%*

4.9. táblázat. A megfigyelt kárgyakoriságok eloszlása

Ezek alapján lehetséges, hogy nemcsak azért illeszkedik jobban a 4. modell az adatokra, mint az 1. és a 2. modell, mert a kárgyakoriság jobban magyarázza az átlagkárt, mint a kárdarabszám, hanem mert több megfigyelés is van a nagyobb kárgyakoriság intervallumokban, mint ahány megfigyelés volt a nagyobb kárdarabszámok esetén.

Összességében elmondható, hogy az aggregált károk modellezése során az átlagkár meghatározásakor sokkal célszerűbb a kárgyakoriságot tekinteni magyarázó változóként, mint az [1] és [2] irodalmakban is alkalmazott kárdarabszámot.

5. Modell

Ebben a modellben a következő változót tettem az átlagkár magyarázó változói közé: log(`kargyakorisag`). Ennek a változónak a p-értéke a program szerint $5,43 \cdot 10^{-14}$, amely még jobb, mint a 4. modell esetén, továbbá itt a modell eltérése a teljes modelltől 5 673, amely kisebb, mint a 4. modell eltérése. Ezek alapján az 5. modell jobban illeszkedik az adatokra, mint a korábbi modellek, így érdemes részletesebben is megvizsgálni.

A modell a következő módon becsüli az i-edik szerződő átlagkárának várható értékét:

$$\hat{\mu}_i = \texttt{kargyakorisag}^{\beta_6} \cdot \exp\big\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1\big\}.$$

A becsült paramétereket pedig az alábbi táblázat tartalmazza:

β_0	7,10582
β_1	0,17053
β_2	-0,64209
β_3	0,17587
β_4	0,35464
β_5	0,24820
β_6	0,30385

4.10. táblázat. Az 5. modell esetén becsült paraméterek

Láthatjuk, hogy az eddigi modellek közül itt a legkisebb az alaposztály becslése. Fontos azonban megjegyezni, hogy korábban, amikor a kárdarabszám került a magyarázó változók közé, akkor az alaposztályba azon megfigyelések tartoztak, akiknek 1 volt a kárdarabszáma, viszont abban az esetben, ha a kárgyakoriságot választjuk magyarázó változónak, akkor az alaposztályba azok tartoznak, akiknek 1 volt a kárgyakorisága (azaz 1 év alatt 1 kárt okoztak).

A 4.10 táblázat alapján láthatjuk, hogy egy alaposztálybeli szerződő átlagos kárnagysága (amennyiben 1 év alatt 1 kárt okoz) várhatóan ebben a modellben $1^{0,30385} \cdot e^{7,10582} \approx 1\ 219$ \$, míg a 4. modell esetén $e^{7,26733+0,01975} \approx 1\ 461$ \$. Mivel az alaposztály az a szegmens, amely a legtöbb megfigyelést tartalmazza, így ez azt jelenti, hogy a 4. modell a legtöbb megfigyelésre túlbecsülte az átlagos kárnagyság várható értékét. Ezt azzal lehet magyarázni, hogy a 4. modell minden megfigyelésről azt feltételezte, hogy 1 évet töltött kockázatban. Láthattuk viszont, hogy amennyiben egy megfigyelésnek kevesebb a kockázatban töltött ideje, akkor az átlagkára is kisebb, és emiatt alacsonyabb lesz az átlagkár becslése, ha a kockázatban töltött időt is figyelembe vesszük a modellezés során. Ebből persze még nem következik, hogy az 5. modell nem becsüli túl az átlagos kárnagyságot, azonban jobb becslést ad, mint a 4. modell.

Vizsgáljuk most meg a kárgyakoriság hatását az átlagkárra ebben a modellben. Az egyszerűség kedvéért tekintsünk olyan szerződőket, akik az alaposztályban vannak, továbbá a kockázatban töltött idejük 1, így a kárgyakoriság megegyezik a kárdarabszámmal. Ha egy szerződő 1 kárt okozott, akkor az ő várható átlagkára $1^{0,30385} \cdot e^{7,10582} \approx 1\ 219$ \$; ha 2 kárt okozott, akkor $2^{0,30385} \cdot e^{7,10582} \approx 1\ 505$ \$; ha 3 kárt okozott, akkor pedig $3^{0,30385} \cdot e^{7,10582} \approx 1\ 702$ \$. Láthatjuk, hogy a becsült β_6 paraméterrel az átlagkár értéke jelentősen növekszik, ha növeljük a kárgyakoriságot.

Mivel ez a modell bizonyult legjobbnak a vizsgált modellek közül, így ezt fogom alkalmazni az aggregált károk modelljénél.

4.3.3. Aggregált károk modellje

Fontos, hogy mivel az 5. modell bizonyult a legjobbnak a vizsgált átlagkár modellek közül, amely nem a kárdarabszámot, hanem a kárgyakoriság logaritmusát tartalmazza magyarázó változóként, így nem lehet alkalmazni a (3.3) egyenlet aggregált károkra vonatkozó becslését. Így először meg kell vizsgálni, hogy hogyan alakul a becslés a jelenlegi esetben.

Jelölje az i-edik szerződő kockázatban töltött idejét ω_i. Ekkor a 3. fejezet jelöléseit alkalmazva az i-edik szerződő átlagkárának várható értéke a következőképpen néz ki, amennyiben a magyarázó változók között a kárdarabszám (N_i) helyett a kárgyakoriság logaritmusa ($\log(N_i/\omega_i)$) szerepel:

$$\mu_i^{(N)} = \mathbb{E}(\overline{Y_i} \mid N_i, x_i) = e^{x_i\beta+\beta_N \log\frac{N_i}{\omega_i}} = e^{x_i\beta}\left(\frac{N_i}{\omega_i}\right)^{\beta_N} = \mu_i\left(\frac{N_i}{\omega_i}\right)^{\beta_N}.$$

Fontos megjegyezni, hogy μ_i olyan alakú, mint a független esetben az átlagkár becslése, viszont az összefüggő eset β paramétervektorát tartalmazza. Ezek alapján az aggregált károk várható értéke a következő:

$$\begin{aligned}\mathbb{E}(S_i \mid x_i) &= \mathbb{E}\left(N_i \cdot \mathbb{E}(\overline{Y_i} \mid x_i, N_i) \mid x_i\right) = \mathbb{E}\left(N_i\mu_i\left(\tfrac{N_i}{\omega_i}\right)^{\beta_N} \mid x_i\right) \\ &= \mu_i \cdot \frac{1}{\omega_i^{\beta_N}} \cdot \mathbb{E}\left(N_i^{1+\beta_N}\right).\end{aligned}$$

Ebben az esetben tehát egészen másképp kell kiszámolni az aggregált károk várható értékét, mint ahogy azt korábban az elmélet bemutatása során láthattuk: közvetlenül nem kell felhasználni a kárdarabszámra vonatkozó becslést ($\hat{\nu}_i$), az csak közvetetten kerül a modellbe N_i-n keresztül, hiszen $N_i \sim Poisson(\hat{\nu}_i)$. Bár ez nem olyan szép akadémikus eredmény, mint abban az esetben, amikor a kárdarabszámot tekintjük magyarázó változónak, de ahogy láthattuk, ennek a gyakorlati haszna sokkal nagyobb.

Jelen esetben β_N az 5. modellben szereplő `kargyakorisag` változóhoz tartozó becsült paraméter, azaz $\beta_N = \beta_6 = 0,30385$, és $\mu_i = \exp\{\beta_0 + x_1\beta_1 + \ldots + x_5\beta_5\}$, ahol β_0,...,β_5 szintén az 5. modell becsült paraméterei.

Ahhoz, hogy ki tudjam számolni az aggregált károk várható értékének becslését, először minden i szerződőre az $N_i^{1,30385}$ várható értékét kellett megbecsülnöm. Mivel erre nincs egzakt formula, így ezt szintén az `R` programban valósítottam meg úgy, hogy minden i szerződőre szimuláltam 100 000 adatot, melyeknek eloszlása $Poisson(\hat{\nu}_i)$, a kapott értékeket az 1,30385. hatványra emeltem, majd kiátlagoltam ezeket az értékeket. A szimuláció során leteszteltem, hogy elegendő-e a 100 000 elemszámú minta használata, és azt az eredményt kaptam, hogy igen, ugyanis a szimuláció megismétlése folyamán ugyanazt az eredményt kaptam 2 tizedesjegy pontossággal.

Fontos, hogy az aggregált károk várható értékének becslésére kapott képletben μ_i nem az átlagkár becsült várható értéke, hiszen ebben nem szerepel

a `kargyakorisag` változó és a hozzá tartozó β_6 paraméter. Ezáltal nem használhattam a program által adott átlagkár becsléseket (mert azok már a kárgyakoriság hatását is tartalmazzák), így ezt a μ_i értéket is külön kiszámoltam minden i szerződőre.

Ezek után a teszt adatokon becsült összkárok és a valódi összkárok átlagos abszolút eltérésére a következő eredményt kaptam:

$$\sum_{i=1}^{13\ 735} \frac{|S_i - \mu_i \cdot \omega_i^{-1,30385} \cdot \mathbb{E}\left(N_i^{1,30385}\right)|}{13\ 735} = 246,7495;$$

az átlagos négyzetes eltérésre pedig:

$$\sum_{i=1}^{13\ 735} \frac{(S_i - \mu_i \cdot \omega_i^{-1,30385} \cdot \mathbb{E}\left(N_i^{1,30385}\right))^2}{13\ 735} = 1\ 081\ 020.$$

4.4. Az eredmények összehasonlítása

Ebben a fejezetben bemutatom, hogy milyen eredmények jöttek ki az aggregált károk modelljére, amikor függetlenséget, illetve amikor összefüggést feltételeztem a kárdarabszám és az átlagkár között. Az összefüggő eset átlagkár modelljének az 5. modellt választottam, mivel ez bizonyult a legjobbnak a vizsgált modellek közül. Az összehasonlításokhoz érdemes felidézni, hogy az i-edik szerződő aggregált kárjainak várható értéke a független esetben

$$\mathbb{E}(S_i \mid x_i) = \hat{\nu}_i \cdot \hat{\mu}_i,$$

ahol $\hat{\nu}_i$ a kárdarabszám várható értékének becslése, $\hat{\mu}_i$ pedig az átlagkáré; míg összefüggő esetben

$$\mathbb{E}(S_i \mid x_i) = \mu_i \cdot \frac{1}{\omega_i^{\beta_N}} \cdot \mathbb{E}\left(N_i^{1+\beta_N}\right),$$

ahol ω_i jelöli a kockázatban töltött időt, N_i a kárdarabszám, β_N a kárgyakorisághoz tartozó paraméter, és fontos megjegyezni, hogy $\mu_i = e^{x_i\beta}$ nem tartalmazza a kárgyakoriság változót.

A kárdarabszám várható értékének becslése megegyezik a független és az összefüggő esetben, tehát vizsgáljuk meg először, hogy milyen paraméterbecsléseket kaptunk az átlagkár modellben a független és az összefüggő esetben. Emlékeztetőül az átlagkár modellje a következő:

$$\hat{\mu}_i = \texttt{kargyakorisag}^{\beta_6} \cdot \exp\big\{\beta_0 + \beta_1 \cdot \texttt{gepj\_tipus}_A + \beta_2 \cdot \texttt{gepj\_tipus}_B + \beta_3 \cdot \texttt{nem}_{ferfi} + \beta_4 \cdot \texttt{lakhely}_A + \beta_5 \cdot \texttt{kor}_1\big\}.$$

Ez a képlet az összefüggő modell esetén szerepelt, de teljesül a független esetre is, $\beta_6 = 0$ választással. A független és az összefüggő eset paramétereinek

Paraméter	Független modell	Összefüggő modell
$\exp(\beta_0)$	1 545,511	1 219,041
$\exp(\beta_1)$	1,170	1,186
$\exp(\beta_2)$	0,505	0,526
$\exp(\beta_3)$	1,197	1,192
$\exp(\beta_4)$	1,492	1,426
$\exp(\beta_5)$	1,338	1,282

4.11. táblázat. Az átlagkár modellek becsült paraméterei

összehasonlítását segíti a következő táblázat, amely az $\exp(\beta_i)$-ket tartalmazza 3 tizedesjegyre kerekítve.
Mivel az összefüggő esetben $\beta_6 = 0,30385$ az átlagkár várható értéke becslésének képletében nem az exponenciális függvényben szerepel, így ennek az együtthatónak a hatását külön kell vizsgálni. A táblázatban láthatjuk, hogy az alaposztály átlagkárának becslése a független modell esetén kb. 1 546 \$, míg az összefüggő modell esetén a kárgyakoriságtól függően változik, azaz például ha a szerződő kárgyakorisága 1, akkor az átlagkár becslése kb. 1 219 \$, de ha például a szerződő kárgyakorisága 3, akkor az átlagkár becslése kb. $1\,219 \cdot 3^{0,30385} \approx 1\,702$ \$. Tudjuk, hogy az összefüggő modell jobb, hiszen kisebb a távolsága a teljes modelltől, mint a független modellé. Ezek alapján kisebb kárgyakoriságok esetén a független modell túlbecsüli az átlagkár várható értékét, míg nagyobb kárgyakoriságok esetén alulbecsüli.

Vizsgáljuk most meg, hogy melyik modell illeszkedik jobban a tesztadatokra. Az alábbi táblázat összefoglalja a független és az összefüggő eset átlagos abszolút és négyzetes eltéréseit.

	Független modell	Összefüggő modell
Átlagos abszolút eltérés	250,8657	246,7495
Átlagos négyzetes eltérés	1 083 092	1 081 020

4.12. táblázat. A becslések eltérései a tesztadatok értékeitől

Láthatjuk, hogy az összefüggő esetben mind az átlagos abszolút, mind az átlagos négyzetes eltérés kisebb. Ezek alapján arra lehet következtetni, hogy határozottan jobb modellt kapunk abban az esetben, ha összefüggést feltételezünk az átlagos kárnagyság és a kárgyakoriság között, mint ha függetlenséget teszünk fel.

A [2] irodalom alapján egy további mutatóval is megvizsgálom a modellek közti eltérést, ez az *átlagos százalékos eltérés* (angolul *average percent difference*, röviden APD). A [2] irodalomban ez a mutató az átlagkár és a kárdarabszám közti összefüggéstől függött (mivel ott a kárdarabszámmal magyarázták az átlagkárt), az én esetemben viszont ez a mutató az átlagkár és a kárgyakoriság logaritmusa közti összefüggéstől, azaz a β_N paramétertől függ. Jelölje most a független modell esetén a kárdarabszám és az átlagkár becslését $\hat{\nu}_i$ és $\hat{\mu}_i$, az összefüggő modell

esetén pedig (ahogy korábban is) legyen μ_i az az érték, amelyet úgy kapunk, hogy vesszük az összefüggő modell átlagkárának becslését úgy, hogy β_N helyére nullát helyettesítünk (azaz a független modell átlagkár becslésének képletébe az összefüggő modell paraméterbecsléseit helyettesítjük). Ekkor, amennyiben a megfigyelések száma m, akkor az átlagos százalékos eltérést a következő módon lehet kiszámítani:

$$\mathrm{APD}(\beta_N) = 100 \cdot \frac{1}{m} \sum_{i=1}^{m} \mathrm{APD}_i(\beta_N),$$

ahol

$$\mathrm{APD}_i(\beta_N) = \frac{\mu_i \cdot \omega_i^{-\beta_N} \cdot \mathbb{E}\left(N_i^{1+\beta_N}\right)}{\hat{\nu}_i \hat{\mu}_i} - 1.$$

Ez az átlagos százalékos eltérés tehát megmutatja, hogy az összefüggő esetben az aggregált károk modelljének becslése mennyire tér el a független eset becslésétől. Jelen esetben $m = 13\ 735$ a tesztadatok száma, $\beta_N = \beta_6 = 0,30385$, és így az átlagos százalékos eltérés 17,3137. Ez azt jelenti, hogy az összefüggő modell becslései átlagosan kb. 17 százalékkal nagyobbak, mint a független modell becslései. Fontos azonban azt is figyelembe venni, hogy $\min(\mathrm{APD}_i) = -24,8657$ és $\max(\mathrm{APD}_i) = 549,718$. Tehát az is előfordul, hogy az összefüggő modell becslése kb. 25%-kal kisebb, mint a független esetben, a legnagyobb eltérésnél pedig az összefüggő modell becslése kb. 550%-kal nagyobb. Ebből is láthatjuk, hogy a kárgyakoriság figyelembevételével teljesen megváltozik az átlagos kárnagyság becslése, ráadásul a független esetben kb. 17%-kal alacsonyabb átlagkár jönne ki a portfólióra, amely egy díjkalkuláció során jelentős díjhiányhoz vezethet.

4.4.1. Az illeszkedések vizsgálata

Az 1.1.5. fejezetben leírtak alapján most megvizsgálom, hogy az egyes modellek mennyire illeszkednek jól az adatokra. Ehhez azt fogom felhasználni, hogy amennyiben a modellben a megfigyelések száma m, a becsült paraméterek száma pedig p, és a modell távolsága a teljes modelltől $D(y, \hat{\mu})$, akkor

$$\frac{D(y, \hat{\mu})}{\phi} \sim \chi^2_{m-p}.$$

Az alábbiakban megvizsgálom, hogy hogyan illeszkedik az adatokra a kárszám modell, az átlagkár modell a független és az összefüggő esetben, továbbá összehasonlítom egymással az utóbbi két modellt, hiszen ezek egymásba ágyazott modellek.

Kárdarabszám modell

A kárdarabszám modell esetén azon megfigyeléseket vettem figyelembe, amelyeknek az adattípusa (az általam készített `datatype` változóban) „training”,

és a kockázatban töltött idejük nagyobb, mint 0. Az ilyen megfigyelések száma 54 121, továbbá ebben a modellben a becsült paraméterek száma $p = 11$, a modell távolsága a teljes modelltől pedig $D(y, \hat{\nu}) = 20\ 185$. A program becslése a szórásparaméterre ebben a modellben $\phi = 1$, így ezek alapján a tesztstatisztika értéke

$$\frac{D(y, \hat{\nu})}{\phi} = \frac{20\ 185}{1} = 20\ 185.$$

A $\chi^2_{54\ 110}$ eloszlás 95. percentilise (egészre kerekítve) 54 652, tehát a tesztstatisztika értéke kisebb, mint a kritikus érték, ami azt jelenti, hogy a modell jól illeszkedik az adatokra.

Átlagkár modell, független eset

A független átlagkár modellezés során azokat az adatokat vettem figyelembe, amelyeknek az adattípusa „training", és a kárszámuk legalább 1 volt. Ezen megfigyelések száma $m = 3\ 688$, és jelen esetben $p = 6$, továbbá $D(y, \hat{\mu}) = 5\ 860$. A program a szórásparaméterre $\phi = 3,21409$ becslést adott, így tehát

$$\frac{D(y, \hat{\mu})}{\phi} = \frac{5\ 860}{3,21409} \approx 1\ 823,22.$$

Továbbá a $\chi^2_{3\ 682}$ eloszlás 95. percentilise (egészre kerekítve) 3 824, tehát a tesztstatisztika értéke jóval kisebb, mint a kritikus érték, azaz a modell jól illeszkedik az adatokra.

Átlagkár modell, összefüggő eset

Ezt a modellt is ugyanazon adatok alapján készítettem, mint a független esetben. Tehát ebben az esetben szintén $m = 3\ 688$, viszont $p = 7$, továbbá $D(y, \tilde{\mu}) = 5\ 673$, és a program jelen esetben $\phi = 2,825123$ szórásparaméterrel számolt, így:

$$\frac{D(y, \tilde{\mu})}{\phi} = \frac{5\ 673}{2,825123} \approx 2\ 008,054.$$

A $\chi^2_{3\ 681}$ eloszlás 95. percentilise (egészre kerekítve) 3 823, tehát ez a modell is jól illeszkedik az adatokra.

Aggregált károk modellje

Vizsgáljuk most meg egymásba ágyazott modellek összehasonlításával, hogy a független modell használható-e az összefüggő modell helyett. Ehhez elég az átlagkár modelleket összehasonlítani. A nullhipotézisünk az, hogy az összefüggő modellben $\beta_6 = 0$, amellyel tehát visszakapjuk a független modellt. Jelölje továbbra is a független modell távolságát a teljes modelltől $D(y, \hat{\mu})$, az összefüggő modellét pedig $D(y, \tilde{\mu})$, továbbá legyen a független modell becsült paramétereinek száma p_1, az

összefüggő modell paramétereinek száma pedig p_2. Az összehasonlításhoz az (1.9) összefüggést fogom használni, amely alapján

$$\frac{D(y,\hat{\mu}) - D(y,\tilde{\mu})}{\phi} \sim \chi^2_{p_2-p_1}.$$

Mivel az összefüggő eset átlagkár modellje a bővebb modell, így ennek a modellnek a becsült szórásparaméterével kell számolni, hiszen ha nem teljesül a nullhipotézis, akkor ez a becslés a helytálló. Jelen esetben tehát $\phi = 2,825123$, így ha behelyettesítjük a megfelelő értékeket, akkor a következőt kapjuk:

$$\frac{D(y,\hat{\mu}) - D(y,\tilde{\mu})}{\phi} = \frac{5\ 860 - 5\ 673}{2,825123} = \frac{187}{2,825123} \approx 66,19.$$

Jelen esetben p_1=6 és p_2=7, tehát a χ^2_1 eloszlás 95. percentilisét kell tekinteni, amely (2 tizedesjegyre kerekítve) 3,84. Így tehát azt kaptuk, hogy a tesztstatisztika értéke jóval nagyobb, mint a kritikus érték, tehát elutasítjuk a nullhipotézist, miszerint $\beta_6 = 0$. Ez pedig azt jelenti, hogy a független modell nem használható az összefüggő modell helyett.

5. fejezet

Összefoglalás

5.1. Megállapítások, eredmények

Szakdolgozatom célja az volt, hogy megvizsgáljam az általánosított lineáris modell segítségével, hogy van-e összefüggés a kárdarabszámok és az átlagos kárnagyságok között. Ahogy láthattuk, arra az eredményre jutottam, hogy ez a feltevés nem teljesen állja meg a helyét (bár ekkor is jobb illeszkedést kapunk, mint ha függetlenséget feltételeznénk), hiszen ha azt tesszük fel, hogy a kárgyakoriság és az átlagkár között áll fenn összefüggés, akkor egy sokkal jobb modellt kapunk eredményül.

A modellezések eredményeiből megállapítható, hogy amennyiben egy szerződő aggregált kárainak a várható értékét szeretnénk becsülni az általánosított lineáris modell segítségével, és függetlenség helyett összefüggést feltételezünk a kárgyakoriság és az átlagos kárnagyság között, akkor szignifikáns eltérést kapunk, hiszen:

- az átlagkár modellezése során az összefüggő esetben a kárgyakoriság logaritmusának becsült együtthatója ($\beta_6 = 0,30385$) azt mutatja, hogy elég erős összefüggés áll fenn a két változó között (minél nagyobb egy szerződő kárgyakorisága, annál nagyobb lesz a várható átlagkára);
- az összefüggő modell esetén a tesztadatokon a várható és a valódi összkárok átlagos abszolút és négyzetes eltérése is kisebb, mint a független modell esetén;
- a független és az összefüggő modell átlagos százalékos eltérése jelentős: az összefüggő modell esetén az aggregált károk várható értéke átlagosan 17%-kal nagyobb, sőt, az eltérés akár 550%-os is lehet, ráadásul az is előfordul, hogy az összefüggő modell becslése 25%-kal kisebb, mint a független modellé (tehát az eltérés előjele nem egyértelmű);

- ha egymásba ágyazott modellként vizsgáljuk a független és az összefüggő aggregált kármodellt, akkor nem állítható, hogy a független modell egy alkalmas helyettesítése lenne az összefüggő modellnek.

Ezek az eredmények mind arra utalnak, hogy az a modell, amelyben összefüggést feltételezünk, jobbnak bizonyul a független modellnél. Nagyon fontos, hogy illeszkedésvizsgálat alapján az összefüggő esetben az átlagkár modell jól illeszkedik az adatokra, ráadásul a tesztadatokra is jobban jelezte előre az átlagkárokat, mint a független eset átlagkár modellje. Vegyük azonban figyelembe, hogy a kárdarabszámokra vonatkozó Poisson eloszlás feltételezése – ahogyan azt korábban láthattuk – nem biztos, hogy megállja a helyét, és lehet, hogy egy kevert eloszlás alkalmazása helyesebb döntés lenne.

Fontos megjegyezni, hogy szakdolgozatomban csak egyetlen év szerződéseit vizsgáltam, így csak úgy tudtam a modell illeszkedését vizsgálni, hogy felosztottam a rendelkezésre álló adatokat: 80%-ukra becsültem a modellt, és 20%-ukra illesztettem, majd megvizsgáltam, hogy a tesztadatokon az előrejelzések mennyire térnek el a valós értékektől. Megbízhatóbb eredményeket kaphatnánk, amennyiben több évnyi megfigyelés állna rendelkezésre, így több év adatai alapján lehetne becsülni a modellt, és a maradékra pedig vizsgálni az előrejelzés megfelelőségét. Továbbá, figyelembe kell venni azt is, hogy én csak gépjármű biztosítások szerződéseit vizsgáltam, így más típusú biztosítások (pl. lakásbiztosítás) esetén egyáltalán nem biztos, hogy bármilyen összefüggés is fennáll a kárdarabszám és az átlagos kárnagyság között.

Összességében úgy gondolom, hogy nem érdemes feltételezni a kárszámok és kárnagyságok között összefüggést, azonban nagyon is érdemes azt feltenni, hogy a kárgyakoriságok és kárnagyságok között áll fenn összefüggés. Annak érdekében azonban, hogy pontosabb képet kapjunk, fontos, hogy több szempontból is megvizsgáljuk ezt a témakört, hiszen egyetlen adathalmazon végzett modellezés alapján még nem lehet kimondani, hogy mindig összefüggés áll fenn ezen két változó között.

5.2. További modellezési lehetőségek

Ahhoz, hogy általánosságban beszélhessünk a kárgyakoriság és az átlagkár közti összefüggésről, az a legfontosabb, hogy többféle biztosítás esetén is modellezzünk. Amennyiben azonban maradunk a gépjármű biztosítások eseténél, szintén vannak további lehetőségek, hogy hogyan vizsgáljuk az összefüggést (amelyeket természetesen alkalmazhatunk más típusú biztosítások esetén is).

Ahhoz, hogy a lehető legpontosabb képet kapjuk, fontos, hogy a modellezés során a megfelelő eloszlásokat válasszuk. Mivel a biztosításban leggyakrabban Poisson-eloszlásúnak feltételezik a kárdarabszámot, és Gamma eloszlásúnak az átlagkárt, így én is ezen eloszlásokkal dolgoztam. Lehetséges azonban, hogy nem ezek a legmegfelelőbb eloszlások egy adott adathalmazra.

További lehetőség az aggregált károk modellezésére az általánosított lineáris modellek körében az ún. *Tweedie modellek* alkalmazása, amelyek szintén az exponenciális szórásmodellhez tartoznak. Ebben az esetben közvetlenül az összkárt lehet modellezni úgy, hogy azt Tweedie eloszlásúnak feltételezzük. Erről az eljárásól bővebben olvashatunk a következő irodalmakban: [1], [3], [4] és [6].

Egy másik megközelítése a modellezésnek, amikor illesztünk egy-egy általánosított lineáris modellt a kárgyakoriságra és a kárnagyságra is, és ezeket egy kopula segítségével kötjük össze. Amennyiben így szerenénk megvizsgálni a kárgyakoriságok és kárnagyságok közti összefüggést, akkor egy jó kiindulás lehet a [7] illetve a [8] irodalom.

Összességében elmondható, hogy a szakdolgozatomban kifejtett modellek többféle módon is tovább fejleszthetők annak érdekében, hogy a kárnagyságok és az átlagkárok közötti összefüggés feltételezésével minél pontosabban meg tudjuk becsülni egy szerződő kárszükségletét. A szakdolgozatban elvégzett modellezési vizsgálatok alapján azonban mindenképpen megállapítható, hogy a kárgyakoriságok és a kárnagyságok közötti kapcsolat feltételezése a valóságot jobban megközelítő becsléseket eredményez, mint ha a kárszámok és a kárnagyságok között feltételeznénk összefüggést, vagy ha egyáltalán nem tennénk fel, hogy fennáll bármilyen kapcsolat ezen változók között.

Irodalomjegyzék

[1] J. Schulz: *Generalized linear models for a dependent aggregate claims model (Master's thesis)*, Concordia University, Montréal, Canada, 2013, p. 106.
https://spectrum.library.concordia.ca/977691/1/Schulz_MSc_F2013.pdf

[2] J. Garrido, C. Genest, J. Schulz: *Generalized linear models for dependent frequency and severity of insurance claims*, Insurance: Mathematics and Economics, ISSN 0167-6687, Vol. 70, 2016, pp. 205-215.
http://dx.doi.org/10.1016/j.insmatheco.2016.06.006

[3] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi: *A Practitioner's Guide to Generalized Linear Models*, Watson Wyatt, 2007, p. 113.
https://www.towerswatson.com/DownloadMedia.aspx?media={E7F1DAFE-D085-4169-81CE-C22ED018FBA3}

[4] E. Ohlsson, B. Johansson: *Non-Life Insurance Pricing with Generalized Linear Models*, Springer, Berlin, Heidelberg, ISBN 978-3-642-10790-7, 2010, p. 133.
https://doi.org/10.1007/978-3-642-10791-7

[5] V. Prokaj: *Általánosított lineáris modell (GLM)*, egyetemi jegyzet, 2017. pp. 1-11.
https://prokajvilmos.web.elte.hu/16-17ii/biztmat/GLM.pdf

[6] P. De Jong, G. Z. Heller: *Generalized linear models for insurance data*, Cambridge University Press, ISBN 978-0-521-87914-9, 2008, p. 195.
http://www.acst.mq.edu.au/GLMsforInsuranceData

[7] N. Krämer, E. C. Brechmann, D. Silvestrini, C. Czado: *Total loss estimation using copula-based regression models*, Insurance: Mathematics and Economics, ISSN 0167-6687, Vol. 53. (3), 2013, pp. 829-839.
http://www.sciencedirect.com/science/article/pii/S0167668713001364

[8] C. Czado, R. Kastenmeier, E. C. Brechmann: A. Min: *A mixed copula model for insurance claims and claim sizes*, Scandinavian Actuarial Journal, Online ISSN: 1651-2030, 2012, pp. 278-305.
https://doi.org/10.1080/03461238.2010.546147

Nyilatkozat

Név: Maros Alexandra
ELTE Természettudományi Kar, szak: Biztosítási- és pénzügyi matematika MSc.
NEPTUN azonosító: N9T3HS
Szakdolgozat címe: Kárszámok és kárnagyságok közti kapcsolat modellezése

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