Eötvös Loránd Tudományegyetem

Természettudományi Kar

Hegel Patrik

Biztosítási kárszámok és károk összefüggése

BSc Szakdolgozat

Témavezető:

Arató Miklós

Valószínűségelméleti és Statisztika Tanszék

Budapest, 2018

Tartalomjegyzék

1. fejezet

Bevezetés

A biztosító egy szerződés megkötésekor arra törekszik, hogy a szerződő számára megfelelő biztosítási díjat határozzon meg. Ez az alapján történik, hogy a szerződőt a jövőben várhatóan mekkora kár éri. A gyakorlatban a biztosító a károkon kívül egyéb költségeket is figyelembe vesz, de a szakdolgozatban csak az összkár modellezésével foglalkozom.

A nem-életbiztosítások esetén az összkár modellezése általában úgy történik, hogy a szerződő kárainak a számát és a nagyságát külön-külön modellezik egymástól függetlenül. Ekkor a várható összkár nagysága a kárszám és a kárnagyság várható értékének a szorzataként áll elő. A szerződő kárszámának és kárnagyságának modellezésekor a szerződő tulajdonságai alapján végeznek becsléseket. Ilyen tulajdonság lehet például a vezető életkora, neme, járműje típusa, kora, értéke. A számításoknál gyakran az általánosított lineáris modellt (GLM) alkalmazzák, ami a klasszikus lineáris modell egy általánosítása. A GLM-ben a válasz várható értékét a jellemzők lineáris kombinációjának a függvényeként (kapcsolati függvény segítségével) kapjuk meg.

A megszokott eljárás az, hogy GLM-et illesztünk a kárszámra és az átlagos kárnagyságra. A kárszám és az átlagos kárnagyság várható értékének a szorzataként megkapjuk az összkár várható értékét. Ebben a modellben a kárszám és a kárnagyság között függetlenséget tételeztünk fel. A függetlenség feltételezése azonban esetenként helytelen lehet és nem megfelelő modellt ad. Például járműbiztosítás esetén azoknál a sofőröknél, akik nagy kárt szenvednek várhatóan kevesebb káreset jelenik meg, mint akik több kicsit okoznak.

J. Garrido, C. Genest és J. Schulz cikkében [4] a GLM egy kiterjesztését találjuk az összefüggő esetre. Szakdolgozatomban ezt a modellt is bemutatom és részletezem, majd összehasonlítom a független modellel.

A szakdolgozat első két fejezetében a károk nagyságának és a kárszám modellezésére alkalmazott eloszlásokat és azok tulajdonságait ismertetem. A kár eloszlásoknál a hagyományos eloszlásokon túl kitérek a Pareto eloszlásra, annak általánosítására és az extrém érték eloszlásokra

is. Ezek általában a kárnagyság farokeloszlását hivatottak modellezni, amelynek kiemelt szerepe van a biztosítások esetén. Annak érdekében, hogy minél pontosabb legyen a modell, esetlegesen kisebb módosításokat lehet végezni a hagyományos modelleken, amiket az első fejezet végén mutatok be.
A kárszám modellezésére leggyakrabban Poisson eloszlást alkalmaznak. A második fejezetben ismertetek néhány másik eloszlást, amik szintén jó modellt adhatnak. Az első két fejezet alapjául Arató Miklós Nem-élet biztosításmatematika című jegyzete [1] és Klugman, Panjer és Willmot Loss Models: From Data to Decisions [3] könyve szolgált. Önálló munkaként ebben a részben rövidebb bizonyításokat végeztem.

A harmadik fejezet az összkár modellezéséről szól. Az összkár eloszlásának néhány jellemzője kerül bemutatásra [5] alapján. Önálló munkaként itt néhány példát csináltam, amelyekben a kárszám és kárnagyság eloszlásának paramétereit becsültem, ha megfigyeléseink az összkárt és a kárszámot illetően vannak.

Az általánosított lineáris modell leírását és alkalmazásának egyszerű ismertetését a negyedik fejezet tartalmazza. Alapvetően D. Anderson: A Practitioner's Guide to Generalized Linear Models című jegyzete [2] alapján íródott. Mivel a GLM a klasszikus lineáris modell kiterjesztése, ezért először azt ismertetem ebben a fejezetben, és utána térek át részletesebben a GLM jellemzőire. A GLM egy még részletesebb leírása Heller, Jong: Generalized Linear Models for Insurance Data könyvében [6] található.

Az ötödik fejezetben az összkárnak a GLM keretein belül való modellezéséről írok [4] és [5] alapján. Először a károk száma és nagysága között függetlenség van feltételezve, majd részletezem az összefüggő esetet is. Az összefüggés úgy jelenik meg a modellben, hogy amikor az átlagos kárnagyságra illesztünk GLM-et, akkor a magyarázó változók közé bevesszük a kárszámot is. Ekkor a várható összkár már nem feltétlen egyezik meg a várható kárszám és átlagos kárnagyság szorzatával. Megmutatom, hogy ha a kárszám esetén Poisson eloszlást feltételezünk logaritmikus kapcsolati függvénnyel, akkor az összkár várható értéke 3 tag szorzataként kapható meg: a várható kárszám, egy módosított átlagos kárnagyság és egy az összefüggésért felelős korrekciós tag szorzataként.

A hatodik fejezetben járműbalesetekről készült valós adatsoron ([6]) végzem el a modellezést önálló munkaként. Az összefüggő és független modellt is alkalmazom, és megnézem mennyiben különbözött a két modell.

2. fejezet

Káreloszlások

Egy kár esetén elsősorban a kifizetendő pénzösszeg érdekes [1]. Ennek érdekében fontos ismerni az egyes károkra kifizetett összeg eloszlását. Mivel a lehetséges kárértékek száma nagy, ezért többnyire folytonos eloszlásokat célszerű illeszteni. A következő részben összefoglalom a legtöbbet használt káreloszlásokat és azok néhány tulajdonságát. Tartalmilag és felépítésileg ennek a résznek az alapját az [1] (34-43. oldal) és [3] adják, helyenként kisebb kiegészítések fordulnak elő.

2.1. Hagyományos káreloszlások

Az egyik leggyakrabban használt káreloszlás az exponenciális eloszlás:

$$F_X(x) = 1 - e^{-\lambda x} \ (x > 0) \ , \quad f_X(x) = \lambda e^{-\lambda x} \ (x > 0) \ .$$

Előnye, hogy könnyen lehet vele számolni és a paramétere is könnyen becsülhető a tapasztalati középből (hiszen $EX = \lambda$).

Az exponenciális eloszlás örökifjú. Ez azt jelenti, hogy ha a kárnagyságot exponenciális eloszlással modellezzük, akkor tudván, hogy a kár nagysága elér egy d értéket, a d összegen túli kár várható értéke konstans, azaz nem függ d-től. Vagyis ha például a biztosító csak egy adott károsszeg felett fizet (önrész), akkor a káronkénti kifizetés várható nagysága nem változik, csak a megjelenő károk száma.

Egy másik gyakran alkalmazott káreloszlás a lognormális eloszlás. Az X valószínűségi változó lognormális eloszlású ($X \sim logN(\mu, \sigma^2)$), ha a logaritmusa normális eloszlású, azaz $log(X) \sim N(\mu, \sigma^2)$.

Tehát a sűrűségfüggvénye: $f_X(x) = \frac{1}{\sigma x \sqrt{2\pi}} exp\{-\frac{(log(x)-\mu)^2}{2\sigma^2}\} \quad (x > 0)$.

Legyen $Y \sim N(0,1)$, ekkor $X = e^{\mu+\sigma Y}$. Az X k-adik momentuma ekkor az alábbi:

$$E(X^k) = E(e^{k(\mu+\sigma Y)}) = e^{k\mu} E(e^{k\sigma Y}) = e^{k\mu} M_Y(k\sigma),$$

[1] itt nem foglalkozok a kárkifizetési késlekedésekkel

ahol M_Y az Y változó momentum generáló függvénye. Ez ismert: $M_Y(t) = e^{t^2/2}$.
Tehát $E(X^k) = e^{k\mu+\frac{1}{2}k^2\sigma^2}$.
Speciális esetben, $k = 1$ esetén a lognormális eloszlás várható értékét kapjuk: $EX = e^{\mu+\frac{1}{2}\sigma^2}$.
A k-adik momentumokból a szórásnégyzet is könnyedén számolható: $D^2X = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$.

Tegyük fel, hogy ismerjük M tapasztalati közepet és S^2 szórásnégyzetet. Alkalmazzuk a momentum-módszert a lognormális eloszlás paramétereinek becslésére:

$$\begin{cases} e^{\hat{\mu}+\frac{1}{2}\hat{\sigma}^2} = M \\ e^{2\hat{\mu}+\hat{\sigma}^2}(e^{\hat{\sigma}^2} - 1) = S^2 \end{cases} .$$

Ezt az egyenletrendszert $e^{\hat{\mu}}$-re és $e^{\hat{\sigma}^2}$-re megoldva könnyen kapjuk:

$$e^{\hat{\sigma}^2} = 1 + \frac{S^2}{M^2} , \qquad e^{\hat{\mu}} = \frac{M}{\sqrt{1+\frac{S^2}{M^2}}} \qquad .$$

A Gamma eloszlás a biztosításban az előzőeknél valamivel talán ritkábban használt, de hasonlóan fontos káreloszlás. $X \sim \Gamma(\alpha, \theta)$, ha sűrűségfüggvénye $f_X(x) = \frac{(\frac{x}{\theta})^\alpha e^{-\frac{x}{\theta}}}{x\Gamma(\alpha)}, \quad x, \alpha, \theta > 0$,
ahol $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt, \quad \alpha > 0.$
A Γ-függvényre parciális integrálással adódik a következő tulajdonság: $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
Szokás a Gamma eloszlást úgy is paraméterezni, hogy θ helyett $\beta = \frac{1}{\theta}$ paramétert használnak. A paraméterezés általában attól függ, hogy melyiket kényelmesebb használni. A szakdolgozatomban többnyire az első paraméterezést használom, mert ekkor θ skála-paraméter.
Skála-paraméter: tekintsünk egy olyan eloszlást, amelyet megszorozva egy (pozitív) konstanssal továbbra is ugyanaz az eloszlástípus marad (csak esetleg más paraméterekkel). Ha van olyan paramétere az előző eloszlásnak, amit c konstanssal szorozva az új eloszlás is az eredeti eloszlás c-szerese lesz, valamint ha az eredeti eloszlást konstanssal szorozva csak ez a paraméter változik, akkor ez a paraméter skála-paraméter.

A Gamma eloszlás várható értékének és szórásnégyzetének megállapításához számoljuk ki a k-adik momentumait (hasonlóan, ahogy a lognormális eloszlás esetén is tettük):

$E(X^k) = \int_0^\infty x^k \frac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}dx = \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1}e^{-y}}{\Gamma(\alpha)\theta^\alpha}\theta dx = \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k)$ minden $k > 0$ esetén.
Speciálisan, a Gamma eloszlás várható értéke (kihasználva, hogy $\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \alpha$) $EX = \theta\alpha$.
Kis számolással könnyen kapjuk, hogy a szórásnégyzete pedig $D^2X = \theta^2\alpha$.

Megemlítendő a Gamma eloszlás azon tulajdonsága, hogy független, azonos skálaparaméterű Gamma eloszlású változók összege is Gamma eloszlású, vagyis ha $X_1 \sim \Gamma(\alpha_1, \theta), X_2 \sim \Gamma(\alpha_2, \theta), ..., X_n \sim \Gamma(\alpha_n, \theta)$, akkor $X_1 + X_2 + ... + X_n \sim \Gamma(\alpha_1 + \alpha_2 + ... + \alpha_n, \theta)$. A bizonyításhoz lássuk be két független Gamma-változóra (majd indukció). A Gamma eloszlás momentum generáló függvénye $M(t; \alpha, \theta) = \frac{1}{(1-\theta t)^\alpha}$. Kihasználva a függetlenséget,

$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t) = \frac{1}{(1-\theta t)^{\alpha_1}}\frac{1}{(1-\theta t)^{\alpha_2}} = \frac{1}{(1-\theta t)^{\alpha_1+\alpha_2}}$,
amiből következik, hogy $X_1 + X_2 \sim \Gamma(\alpha_1 + \alpha_2, \theta)$.

Speciálisan, ha $X_1, ..., X_n$ λ-paraméterű exponenciális eloszlású, független változók, akkor $X_1 + ... + X_n \sim \Gamma(n, 1/\lambda)$, ugyanis $Exp(\lambda) \sim \Gamma(1, 1/\lambda)$.

2.2. Pareto eloszlás

Népszerű káreloszlás a Pareto-eloszlás is. $X \sim Pareto(\alpha, \beta)$, (ahol $\alpha, \beta > 0$), ha az eloszlásfüggvénye:
$F_X(x) = 1 - (\frac{\beta}{\beta+x})^\alpha \qquad (x > 0)$.
Sűrűségfüggvénye: $f_X(x) = \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}$.
Egy másik gyakori paraméterezése a Pareto eloszlásnak: $Y \sim Pareto(a, c)$, (ahol $a, c > 0$) ha az eloszlásfüggvénye:
$F_X(x) = 1 - (\frac{c}{x})^a \qquad (x > c)$
Világos, hogy ha $X \sim Pareto(\alpha, \beta)$ az első paraméterezés szerint, akkor $X + \beta \sim Pareto(\alpha, \beta)$ a második paraméterezés szerint.
Ez az eloszlás különösen fontos amiatt, hogy az eddigiekkel ellentétben a várható értéke nem mindig véges. Pontosabban a várható értéke csak $\alpha > 1$ esetén, szórása pedig csak $\alpha > 2$ esetén véges. Annak érdekében, hogy ezt igazoljuk, nézzük a Pareto-eloszlás k-adik momentumait:

$E(X^k) = \int_0^\infty x^k \frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}} dx$, $\quad y = x + \beta$ helyettesítéssel
$= \int_0^\infty (y-\beta)^k \frac{\alpha\beta^\alpha}{y^{\alpha+1}} dy = \alpha\beta^\alpha \int_0^\infty \sum_{j=0}^k \binom{k}{j} y^{j-\alpha-1}(-\beta)^{k-j} dy$ $\quad$ minden k egészre.

Ez az integrál akkor véges, ha a szummában y minden kitevője kisebb, mint -1, vagyis $j - \alpha - 1 < -1$ minden j-re, ami ugyanakkor teljesül, ha $k < \alpha$. Tehát speciálisan, a várható érték akkor véges, ha $\alpha > 1$, a szórás(négyzet) pedig akkor véges, ha $\alpha > 2$.

Sokszor a biztosítások esetén a veszteségek legnagyobb részét a nagy összegű károk teszik ki. Ezért fontos a káreloszlások "farokeloszlásait" vizsgálni (azaz a nagy értékek esetén milyen az eloszlás). Azokat az eloszlásokat, amelyek hajlamosabbak nagyobb valószínűséggel felvenni nagy értékeket, nehéz farkúnak ("heavy-tailed") nevezzük. Modellválasztás esetén a farokeloszlás "nehézsége" szűkítheti a lehetséges jól illeszkedő eloszlások körét. Érezzük, hogy ez nem egy túl precízen definiált fogalom. Például a k-adik momentumok létezése/nem létezése minden k-ra jó jelzője lehet a nehéz farkú eloszlásoknak, ugyanis ha az X valószínűségi változó nagy valószínűséggel vesz fel nagy értékeket, akkor $E(X^k) = \int_0^\infty x^k f_X(x)dx$ integrál inkább hajlamos nem véges lenni.
Ez alapján például a Pareto nagy farkú eloszlás, mert csak $k < \alpha$ esetén létezik a k-adik momentuma, ellenben a Gamma eloszlás "kis farkú", hiszen korábban láttuk, hogy $E(X^k) = \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k)$, azaz véges minden $k > 0$ esetén.

2.3. Általánosított Pareto eloszlás

Azonban általában a nagyobb károk ritkán fordulnak elő, ezért kevesebb adat van róluk. Emiatt érdemes lehet érzékenyebb modellt választani a farokeloszlások modellezésére. Egy másik ok lehet, hogy a valós adatok eloszlása bonyolultabb, mint a megszokott modellek. Az Általánosított Pareto eloszlás (GPD) egy gyakran használatos eloszlás ilyen célra.
A GPD egy három-paraméteres eloszlás: μ eltolási paraméter, σ (>0) skála paraméter és ξ alakparaméter esetén ($X \sim GPD(\mu, \sigma, \xi)$ az eloszlásfüggvénye és sűrűségfüggvénye:

$$F_X(x) = \begin{cases} 1 - (1 + \frac{\xi(x-\mu)}{\sigma})^{-1/\xi} & \text{ha } \xi \neq 0 \\ 1 - e^{-\frac{(x-\mu)}{\sigma}} & \text{ha } \xi = 0 \end{cases}$$

$$f_X(x) = \begin{cases} \frac{1}{\sigma}(1 + \frac{\xi(x-\mu)}{\sigma})^{-1-1/\xi} & \text{ha } \xi \neq 0 \\ \frac{1}{\sigma}e^{-\frac{(x-\mu)}{\sigma}} & \text{ha } \xi = 0 \end{cases},$$

ahol $x \geq \mu$ ha $\xi \geq 0$ és $\mu \leq x \leq \mu - \sigma/\xi$ ha $\xi < 0$.
Az eltolási és skála paraméter szerepe egyértelmű. Az alakparaméter előjele aszerint alakulhat, hogy milyen az eloszlás, aminek a farokeloszlását szeretnénk modellezni:

- ha a sűrűségfüggvénye exponenciálisan csökken a farok felé (mint például normális eloszlásnál), akkor ez ahhoz vezet, hogy a GPD alakparamétere 0
- ha a sűrűségfüggvény polinomiálisan csökken (mint például student t eloszlásnál), akkor ez pozitív alakparaméterhez vezet
- ha a sűrűségfüggvény farokrésze véges (mint például beta eloszlás esetén), akkor ez negatív alakparaméterhez vezet.

Jegyezzük meg, hogy ha a GPD ξ alakparamétere és μ eltolási paramétere is 0, akkor az ilyen GPD ekvivalens az exponenciális eloszlással ($1/\sigma$ paraméterű).
Ha pedig az alakparaméter $\xi > 0$ és az eltolási paraméter $\mu = \sigma/\xi$, akkor a GPD ekvivalens a Pareto eloszlással, amelynek a paraméterei (a második paraméterezés szerint) $a = 1/\xi$ és $c = \sigma/\xi$ (vagyis a Pareto valóban speciális esete az általánosított Pareto-nak).
A GPD k-adik momentumaira hasonló állítás igaz, mint a Pareto eloszlás esetén: a GPD k-adik momentuma $k \geq 1/\xi > 0$ esetén végtelen. Várható értéke $\xi < 1$ esetén $EX = \mu + \sigma/(1 - \xi)$.

2.4. Extrém érték eloszlások

Az extrém, ritkán előforduló események bekövetkezésének modellezésével az extrémérték-elmélet foglalkozik. Ilyen események lehetnek például áradások, extrém hőmérsékletek, tőzsdekrach, természeti katasztrófák, stb.

Legyen $X_1, X_2, ...X_n$ független, azonos eloszlású valószínűségi változók sorozata, M_n a maximuma ennek. Az elmélet szerint M_n (aszimptotikus) eloszlásának modellezésére elegendő három eloszlás: a Gumbel, Fréchet és Weibull eloszlások. Definiáljuk ezeket a sűrűségfüggvényük segítségével:

Gumbel eloszlás sűrűségfüggvénye μ eltolási és σ skála paraméterrel:

$f_X(x) = \frac{1}{\sigma}\ exp\{-\frac{x-\mu}{\sigma} - exp\{-\frac{x-\mu}{\sigma}\}\}$

A Gumbel eloszlás értelmezve van az egész valós tengelyen. Alakja nem függ a paraméterektől (nincs alakparaméter).

Fréchet eloszlás sűrűségfüggvénye α (>0) alak és β (>0) skála paraméterrel:

$f_X(x) = \frac{\alpha}{\beta}(\frac{\beta}{x})^{\alpha+1}\ exp\{-(\frac{\beta}{x})^{\alpha}\} \quad (x > 0)$

Tehát csak a pozitív tengelyen van értelmezve, de lehet általánosítani eltolási paraméter (μ) hozzáadásával értelemszerűen. A Fréchet eloszlás másik ismert neve az inverz Weibull eloszlás: ha $X \sim Frechet(\alpha, \beta, \mu = 0)$, akkor $X^{-1} \sim Weibull(\alpha, 1/\beta)$.

Weibull eloszlás sűrűségfüggvénye α (>0) alak és β (>0) skála paraméterrel:

$f_X(x) = \frac{\alpha}{\beta}(\frac{x}{\beta})^{\alpha-1}\ exp\{-(\frac{x}{\beta})^{\alpha}\} \quad (x > 0)$

Hasonlóan a Fréchet eloszláshoz, a Weibull eloszlás esetén is általánosíthatunk eltolási paraméter bevezetésével.

Az extrém érték elmélet foglalkozik azzal, hogy egy adott kezdeti X eloszlás esetén a maximum eloszlása a fenti három eloszlás közül melyiket követi (például a farok nehézsége alapján).

Ezt a három eloszlást kombinálja az általánosított extrémérték eloszlás (Generalized extreme value distribution, GEV) [8]. A GEV eloszlás sűrűségfüggvénye:

$$f_X(x) = \begin{cases} \frac{1}{\sigma}\ exp\{-(1+\xi z)^{-1/\xi}\}\ (1+\xi z)^{-1-1/\xi} & \text{ha } \xi \neq 0 \\ \frac{1}{\sigma}\ exp\{-z - exp(-z)\} & \text{ha } \xi = 0 \end{cases},$$

ahol $z = \frac{x-\mu}{\sigma}$, és ξ, σ (>0), μ rendre alak, skála és eltolási paraméterek.

Az értelmezési tartomány ξ-től függ: ha $\xi = 0$, akkor az eloszlás az egész valós tengelyen értelmezett; ha viszont $\xi \neq 0$, akkor $1 + \xi\frac{x-\mu}{\sigma} > 0$ kell legyen.

Az alakparaméter különböző értékei esetén a GEV megfelel a Gumbel, Fréchet illetve Weibull eloszlásoknak. Ha $\xi = 0$, akkor Gumbel, ha $\xi > 0$, akkor Fréchet, $\xi < 0$ esetén pedig (fordított) Weibull eloszlást kapunk.

2.5. Új eloszlások készítése

A következőekben néhány egyszerű módszert mutatok be, hogy hogyan lehet már ismert eloszlásokból új eloszlásokat készíteni. Példaként az eddig szerepelt eloszlásokkal végzek el ezek közül a transzformációk közül néhányat. Ezeknek a transzformációknak a biztosításban különböző szerepük lehet, azonban néhányat csak a teljesség kedvéért mutatok be.

1. Konstanssal való szorzás

Többek közt az infláció modellezésénél jelenhet meg ez a transzformáció. Ha például az idei év veszteségeit az X valószínűségi változó adja meg, akkor ha 10% az egyenletes infláció a veszteségeken, akkor a jövő évi veszteségeket az $Y = 1.1X$ változó adja meg.
Legyen az X egy folytonos valószínűségi változó, amelynek a sűrűségfüggvénye $f_X(x)$, eloszlásfüggvénye pedig $F_X(x)$. Legyen $Y = \theta X$, ahol $\theta > 0$. Ekkor

$$F_Y(y) = P(\theta X < y) = F_X(y/\theta)$$
$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{\theta}f_X(\frac{y}{\theta}) \ .$$

Ekkor θ paraméter az Y eloszlás skálaparamétere.
Példa: Legyen az X valószínűségi változó eloszlásfüggvénye $F_X(x) = 1 - (1+x)^{-\alpha} \quad (x, \alpha > 0)$. Határozzuk meg az $Y = \theta X$ változó eloszlását!
Az előzőek szerint ekkor $F_Y(y) = 1 - (1 + \frac{y}{\theta})^{-\alpha} = 1 - (\frac{\theta}{\theta+y})^{\alpha}$. Vagyis Y Pareto eloszlású, α és θ paraméterekkel.

2. Hatványra emelés

Legyen X egy abszolút folytonos eloszlású valószínűségi változó, $F_X(x)$ eloszlásfüggvénnyel és $f_X(x)$ sűrűségfüggvénnyel. Az X csak pozitív értékeket vehessen fel, azaz $F_X(0) = 0$. Legyen $Y = X^{1/\tau}$. Ekkor $\tau > 0$ esetén

$$F_Y(y) = P(X < y^{\tau}) = F_X(y^{\tau}) \ , \qquad f_Y(y) = \tau y^{\tau-1} f_X(y^{\tau}) \ , \qquad x > 0 \ ,$$

ha pedig $\tau < 0$, akkor

$$F_Y(y) = P(X \geq y^{\tau}) = 1 - F_X(y^{\tau}) \ , \qquad f_Y(y) = -\tau y^{\tau-1} f_X(y^{\tau}) \ , \qquad x > 0 \ .$$

A sűrűségfüggvényt deriválással kapjuk.
Amikor hatványozunk egy eloszlást, ha $\tau > 0$, akkor a kapott eloszlást transzformáltnak nevezzük, ha $\tau = -1$, akkor inverznek, ha pedig $\tau < 0$ (de $\tau \neq -1$), akkor inverz transzformáltnak nevezzük.
1. Példa: Legyen az X egy Pareto eloszlású változó. Határozzuk meg az eloszlásfüggvényét a transzformált, inverz és inverz transzformált eloszlásnak!
Az eddigieket alkalmazva, ha $Y = X^{1/\tau}$, $\tau > 0$: $\qquad F_Y(y) = F_X(y^{\tau}) = 1 - (\frac{\beta}{\beta+y^{\tau}})^{\alpha} \qquad .$
Ezt az eloszlást Burr-eloszlásnak nevezik. Ha $\tau < 0$, akkor kapjuk az inverz Burr-eloszlást:
$F_Y(y) = 1 - F_X(y^{1/\tau}) = (\frac{\beta}{\beta+y^{\tau}})^{\alpha}$.
Az inverz Pareto eloszlás eloszlásfüggvénye ($\tau = -1$ eset):
$F_Y(y) = 1 - F_X(1/y) = (\frac{\beta}{\beta+\frac{1}{y}})^{\alpha} = (\frac{y}{y+\frac{1}{\beta}})^{\alpha}$.
2. Példa: Legyen az X exponenciális eloszlású változó. Végezzük el rajta az előző három transzformációt! A művelet ugyanaz, mint előbb. A transzformált exponenciális eloszlás megegyezik a Weibull eloszlással (amit korábban már láttunk), az inverz transzformált exponenciális pedig az inverz Weibull eloszlással egyezik meg.

3. Exponenciálás

Legyen az X egy folytonos valószínűségi változó, $F_X(x)$ eloszlásfüggvénnyel és $f_X(x)$ sűrűségfüggvénnyel. Legyen $Y = e^X$. Ekkor $y > 0$ esetén

$$F_X(x) = P(X < log(y)) = F_X(log(y)) \ , \qquad f_X(x) = \frac{1}{y} f_X(log(y)) \qquad .$$

Példa: legyen az X normális eloszlású, μ várható értékkel és σ^2 szórásnégyzettel. Ekkor $Y = e^X$ eloszlása lognormális.

Példa: legyen az X Pareto eloszlású változó α és θ paraméterekkel, és $Y = log(1 + X/\theta)$. Határozzuk meg az Y eloszlását!

$P(Y < y) = P(1 + \frac{x}{\theta} < y) = P(x < \theta e^y - \theta) = 1 - (\frac{\theta}{\theta + \theta e^y - \theta})^\alpha = 1 - e^{-\alpha y}$.

Vagyis Y α paraméterű exponenciális eloszlás. Fordítva, ha Y α paraméterű exponenciális, akkor az $X = \theta e^y + \theta$ változó α és θ paraméterű Pareto eloszlást követ.

4. Keverés

Tekintsünk egy inhomogén közösséget, azaz például a tagjai legyenek besorolhatók több osztályba. Ha a közösség káreseteit szeretnénk vizsgálni, akkor jellemzően a különböző osztályok tagjai különböző károkat okoznak. Az egyes osztályokon belül a tagoktól körülbelül hasonló károkra számítunk. Vagyis például ha egy személy várható kárnagyságát szeretnénk megbecsülni valamilyen eloszlással, akkor előfordulhat, hogy az eloszlás bizonyos paraméterei attól függnek, hogy a személy melyik osztályba sorolható. Azaz tekinthetünk erre úgy, hogy az eloszlás egy paramétere maga is egy valószínűségi változó által felvett érték.

Legyen az X változó sűrűségfüggvénye $f_{X|\Lambda}(x|\lambda)$ és az eloszlásfüggvénye $F_{X|\Lambda}(x|\lambda)$, ahol λ az X egy paramétere (lehet más paramétere is, de az most nem számít). Legyen λ a Λ változó által felvett érték. Ekkor a feltétel nélküli sűrűségfüggvénye X-nek:

$$f_X(x) = \int f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda \ .$$

Az így kapott X eloszlás neve kevert eloszlás. Jegyezzük meg, hogyha Λ diszkrét, akkor az integrál helyett egy szumma áll.

A keverék eloszlások k-adik momentumait a következőképpen kaphatjuk meg (feltételezve, hogy az integrálok felcserélhetőek):

$$E(X^k) = \int \int x^k f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda dx =$$
$$= \int [\int x^k f_{X|\Lambda}(x|\lambda) dx] f_\Lambda(\lambda) d\lambda =$$
$$= \int E(X^k|\lambda) f_\Lambda(\lambda) d\lambda =$$
$$= E[E(X^k|\Lambda)] \ .$$

A tapasztalat azt mutatja, hogy a kevert eloszlások többnyire nehéz farkúak. Ezt mutatom be a következő példán:

Legyen $X|\Lambda$ exponenciális eloszlású, Λ paraméterrel. Legyen Λ gamma eloszlású (α és $\frac{1}{\theta}$ paraméterekkel). Ekkor :

$f_X(x) = \frac{\theta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda e^{-\lambda x} \lambda^{\alpha-1} e^{-\theta\lambda} d\lambda = \frac{\theta^\alpha}{\Gamma(\alpha)} \int_0^\infty \lambda^\alpha e^{-\lambda(x+\theta)} d\lambda = \frac{\theta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(x+\theta)^{\alpha+1}} = \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$.

Ez pedig a Pareto eloszlás sűrűségfüggvénye. Vagyis ha egy exponenciális változó paramétere gamma eloszlást követ, akkor a keverék eloszlás Pareto. A következő példában ezt általánosítom: Most legyen $X|\theta$ gamma eloszlású α és $1/\Theta$ paraméterekkel, továbbá Θ kövessen gamma eloszlást k és $1/\lambda$ paraméterekkel. Ekkor az X eloszlása megegyezik az általánosított Pareto eloszlással:

$f_X(x) = \int_0^\infty f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta =$
$= \int_0^\infty \frac{1}{\Gamma(k)} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{k-1} \theta^{\alpha+k-1} e^{-(x+\lambda)^\theta} d\theta =$
$= \frac{1}{\Gamma(k)} \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{k-1} \frac{\Gamma(\alpha+k)}{(x+\lambda)^{\alpha+k}} \int_0^\infty \frac{1}{\Gamma(\alpha+k)} (x+\lambda)^{\alpha+k} \theta^{\alpha+k-1} e^{-(x+\lambda)^\theta} d\theta =$
$= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(k)} \frac{\lambda^\alpha x^{k-1}}{(x+\lambda)^{\alpha+k}}$

Mivel az utolsó integrál megegyezik egy gamma eloszlás sűrűségfüggvényének integráljával. Ez megegyezik a [3] szerint paraméterezett általánosított Pareto eloszlás sűrűségfüggvényével (α, λ és k paraméterek).

További keverék eloszlásokra példák a [3] és [7]-ben találhatóak.

5. Összeillesztés

Az összeillesztés esetén szintén több eloszlásból kapunk egyet, azzal a különbséggel hogy itt az eloszlás a kár nagyságától függ. Vagyis a modell az, hogy különböző intervallumokon más-más eloszlásokat illesztünk.

Az intervallumok végpontjait jelöljük c_i-kkel, f_i legyen a $[c_{i-1}, c_i]$ intervallumon egy sűrűségfüggvény ($i = 1...k$). Ekkor az összeilleszett eloszlás sűrűségfüggvénye:

$$f_X(x) = \begin{cases} a_1 f_1(x) \ , & c_0 < x < c_1, \\ a_2 f_2(x) \ , & c_1 < x < c_2, \\ \vdots & \\ a_k f_k(x) \ , & c_{k-1} < x < c_k, \end{cases}$$

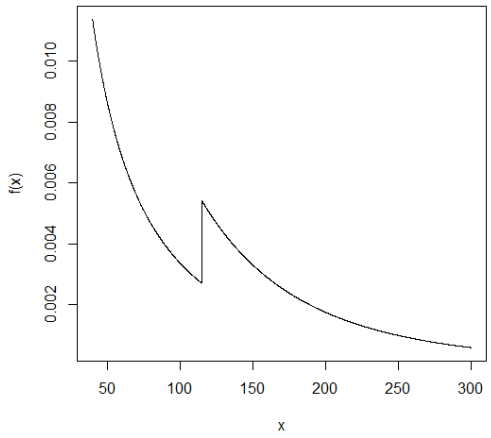
ahol $a_1, a_2, ..., a_k > 0$ konstansok úgy, hogy $a_1 + a_2 + ... + a_k = 1$ (garantálva hogy a kapott függvény sűrűségfüggvény lesz).

Ez a definíció akkor megfelelő, ha a c_i töréspontok előre adottak. Van egy másik lehetőség összeillesztett modellek készítésére, ahol ezeket a töréspontokat könnyebben lehet paraméterként kezelni: legyen $g_i(x)$ egy sűrűségfüggvény és az előző definícióban az $f_i(x)$ helyére $\frac{g_i(x)}{G_i(c_i) - G_i(c_{i-1})}$ -et írjunk, ahol G_i az eloszlásfüggvény.

Egyik oka lehet az összeillesztés alkalmazásának, ha egyes intervallumokon kevesebb adat áll rendelkezésünkre. Ekkor ezeken az értékeken célszerűbb lehet egy az eredetihez képest eltérő eloszlás illesztése vagy empirikus eloszlás alkalmazása. Másik ok lehet a farokeloszlás modellezése: Példa: egy modellben a $[0, c]$ intervallumon exponenciális eloszlást (λ paraméterű), (c, ∞)-on pedig Pareto eloszlást használunk (α, β paraméterű). Ekkor $a_1 = \nu$ és $a_2 = 1 - \nu$ -vel (hogy a teljes integrál 0 legyen)

$$f_X(x) = \begin{cases} \nu \frac{\lambda e^{-\lambda x}}{1-e^{-\lambda x}}, & 0 < x < c, \\ (1-\nu) \frac{\alpha(c+\beta)^\alpha}{(x+\beta)^{\alpha+1}}, & c < x < \infty. \end{cases}$$

A kapott sűrűségfüggvény egyáltalán nem biztos, hogy folytonos. A példában ha például $c = 115, \nu = 0.55, \lambda = 1/110, \beta = 250, \alpha = 4.4$, akkor ahogy a következő ábra is mutatja, nem folytonos.



3. fejezet

Kárszám eloszlások

A károk nagysága mellett fontos ismerni a károk számának eloszlását is, hogy jobban értsük egy biztosítás körülményeit. Így az összesített kárnagyságot is egyszerűbb megbecsülni. A következőkben olyan eloszlásokat mutatok be, amelyek megfelelőek lehetnek a károk számának modellezésére. A kárszám nemnegatív egész szám lehet, így ezek diszkrét eloszlások lesznek. A fejezet alapvetően az [1] (24-33. oldal) és [3]-ra épül.

3.1. Gyakori kárszám eloszlások

A legtöbbször a károk számának modellezésére a Poisson eloszlást alkalmazzák. A Poisson eloszlás használatát igazolja a gyakorlati tapasztalat, de felmerülhet Poisson folyamat eredményeként is. Két fontos tulajdonságát mutatom be az eloszlásnak.
Tegyük fel, hogy a károk száma egy adott időszak alatt Poisson eloszlást követ, és minden kár besorolható m különböző osztály valamelyikébe (pontosan egybe). Például a károkat lehet nagyság szerint osztályozni, vagy baleset esetén a sérülés típusa szerint, stb. Ekkor egy adott osztályba tartozó károk számának eloszlása is Poisson, csak (többnyire) más paraméterrel. Vagyis ha módosítani szeretnénk egy szerződésen (például egy adott típusú kárt nem biztosítani tovább), akkor a kárszám eloszlása továbbra is Poisson marad (természetesen a paramétert újra kell becsülni).
Továbbá nemcsak Poisson eloszlásúak lesznek az egyes osztályok kárszámai, hanem egymástól függetlenek is (azaz például azon károk száma amelyek nagysága egy adott értéknél nagyobb, illetve kisebb, egymástól függetlenek). Ezeket a következő tétel formalizálja:
Legyen N az események száma, $N \sim Poi(\lambda)$. Továbbá minden esemény besorolható m osztály valamelyikébe p_1, p_2, ... , p_m valószínűséggel. Ekkor rendre az osztályokhoz tartozó események száma N_1, N_2, ..., N_m egymástól kölcsönösen független, Poisson eloszlású változók λp_1, λp_2, ... , λp_m paraméterekkel.

Bizonyítás: rögzített $N = n$ -re $(N_1, ..., N_m)$ feltételes együttes eloszlása "multinomiális" $(n, p_1, ...p_m)$ paraméterekkel. Szintén rögzített $N = n$ esetén N_j feltételes peremeloszlása binomiális (n, p_j) paraméterekkel. Vagyis:

$$P(N_1 = n_1, ..., N_m = n_m) = P(N_1 = n_1, ..., N_m = n_m \mid N = n) \times P(N = n) =$$
$$= \frac{n!}{n_1!n_2!...n_m!}\, p_1^{n_1}...p_m^{n_m}\ \frac{e^{-\lambda}\lambda^n}{n!} = \prod_{j=1}^{m} e^{-\lambda p_j}\frac{(\lambda p_j)^{n_j}}{n_j!} \qquad \text{, ahol } n = n_1 + ... + n_m \ .$$

Másrészt N_j peremeloszlása:

$$P(N_j = n_j) = \sum_{n=n_j}^{\infty} P(N_j = n_j \mid N = n)P(N = n) =$$
$$= \sum_{n=n_j}^{\infty} \binom{n}{n_j}\, p_j^{n_j}(1-p_j)^{n-n_j}\ \frac{e^{-\lambda}\lambda^n}{n!} =$$
$$= e^{-\lambda}\frac{(\lambda p_j)^{n_j}}{n_j!}\sum_{n=n_j}^{\infty}\frac{[\lambda(1-p_j)]^{n-n_j}}{(n-n_j)!} =$$
$$= e^{-\lambda}\frac{(\lambda p_j)^{n_j}}{n_j!}e^{\lambda(1-p_j)} =$$
$$= e^{-\lambda p_j}\frac{(\lambda p_j)^{n_j}}{n_j!} \qquad .$$

Vagyis $N_j \sim Poi(\lambda p_j)$ valóban, és tehát mivel az együttes eloszlás a peremeloszlások szorzata, ezért a kölcsönös függetlenség is fennáll.

Például tegyük fel, hogy egy orvosi biztosítás várható kárszáma szerződésenként 3.2 db és a kárszám Poisson eloszlást követ. Vissza akarjuk vonni a fedezetet egy adott ellátásról a szerződésben. Az adatok szerint ez az ellátás az összes káreset körülbelül 5%-át teszi ki. Az előzőek alapján ekkor az új szerződés alatt a károk várható száma $0.95 \times 3.2 = 3.04$. Azonban az összesített kárnagyságról az új szerződés esetén nem igazán lehet mit mondani, ugyanis a visszavont ellátás kárnagyságainak eloszlása különbözhet a többiétől.

A Poisson eloszlás másik hasznos tulajdonsága, hogy ha $N_1, N_2, ..., N_n$ független Poisson eloszlású változók, amelyek paraméterei rendre $\lambda_1, \lambda_2, ..., \lambda_n$, akkor $N = N_1 + N_2 + ... + N_n$ változó is Poisson eloszlású és a paramétere $\lambda = \lambda_1 + \lambda_2 + ... + \lambda_n$.

Ennek belátásához használjuk fel a tényt, hogy független változók összegének generátorfüggvénye az egyes változók generátorfüggvényének szorzata:

$$G_N(z) = \prod_{j=1}^{n} G_{N_j}(z) = \prod_{j=1}^{n} e^{\lambda_j(z-1)} = exp\{\sum_{j=1}^{n}\lambda_j(z-1)\} = e^{\lambda(z-1)} \qquad .$$

Mivel a generátorfüggvény egyértelmű, ezért $N \sim Poi(\lambda)$.

Egy másik gyakran használt kárszám eloszlás a negatív binomiális eloszlás. Mivel két paramétere is van, ezért rugalmasabb a Poisson eloszlásnál. Itt most negatív binomiális eloszlás alatt a "0-ba eltolt" negatív binomiálist értem, azaz ha $N \sim NB(r, q)$, akkor $P(N = k) = \binom{k+r-1}{k}\, q^k(1-q)^r$ (ahol $r > 0$ és $0 \leq q < 1$) .

A várható értéke $\frac{rq}{1-q}$, a szórásnégyzete $\frac{rq}{(1-q)^2}$. Azaz mivel $0 \leq q < 1$, ezért a negatív binomiális eloszlás szórásnégyzete mindig nagyobb mint a várható értéke, ellentétben a Poisson eloszlással, ahol a várható érték megegyezett a szórásnégyzettel. Ez azt jelenti, hogy adott adatok esetén ha a megfigyelt szórásnégyzet nagyobb, mint a várható érték, akkor talán érdemesebb a Poisson eloszlás helyett negatív binomiálist használni a kárszámok modellezésére.

A következőekben megmutatom, hogy a negatív binomiális eloszlás generátorfüggvénye $G(z) = (\frac{1-q}{1-qz})^r$.

Bizonyítás: legyen $p = 1 - q$.

$$G(z) = \sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r q^k z^k = p^r \sum_{k=0}^{\infty} \binom{k+r-1}{k} q^k z^k \ .$$

Vagyis amit meg akarunk mutatni, az nem más, minthogy $y = qz$ jelöléssel:

$$(1-y)^{-r} = \sum_{k=0}^{\infty} \binom{k+r-1}{k} y^k \ .$$

Ehhez fejtsük McLaurin-sorba a $(1-y)^{-r}$ kifejezést, azaz vegyük a Taylor-sorát a 0 körül.

$(1-y)^{-r} = 1 + ry + \frac{1}{2!}(-r)(-r-1)(-y)^2 + \frac{1}{3!}(-r)(-r-1)(-r-2)(-y)^3 + ...$

Nézzük meg y^k együtthatóját:

$\frac{1}{k!}(-r)(-r-1)...(-r-k+1)(-1)^k = \binom{k+r-1}{k}$.

Az imént $y = qz$ helyettesítést használtunk, az állítás innen már következik.

Megjegyzem, hogy a negatív binomiális eloszlás generátorfüggvénye apró algebrai átalakítással a következő alakra hozható:

$G(z) = (1 - \frac{p}{1-p}(z-1))^{-r}$.

A negatív binomiális eloszlás a Poisson eloszlás egyfajta általánosítása: ha a Poisson eloszlás paramétere gamma eloszlást követ, akkor a kapott keverék eloszlás negatív binomiális. A feltételezés, hogy a Poisson eloszlás paramétere valamilyen eloszlást követ (akár diszkrét, akár folytonos) helytálló, ugyanis tekinthetünk a népességre úgy, mintha heterogén lenne, amit a paraméter eloszlása modellez. Például gépjármű biztosítás esetén külön kategóriák alkothatóak a vezető életkora, neme, a jármű által megtett távolság és egyéb vezetési szokások szerint.

Tehát tegyük fel, hogy $N \sim Poi(\lambda)$, ahol λ a Λ változó által felvett érték és $\Lambda \sim Gamma(\alpha, \beta)$. Ekkor N feltételes generátorfüggvénye $G_{N|\Lambda}(z) = e^{\Lambda(z-1)}$. Innen N generátorfüggvénye:

$G_N(z) = E[E(z^N \mid \Lambda)] = E(e^{\Lambda(z-1)}) = L_\Lambda(1-z)$, ahol L_Λ a Laplace-transzformáltja Λ-nak. Ha $\Lambda \sim Gamma(\alpha, \beta)$, akkor a Laplace-transzformáltja $L_\Lambda(1-z) = (\frac{\alpha}{\alpha+(1-z)})^\beta = (\frac{1-\frac{1}{1+\alpha}}{1-\frac{1}{1+\alpha}z})^\beta$, ami egy $NB(\beta, \frac{1}{1+\alpha})$ eloszlású változó generátorfüggvénye.

A η valószínűségi változót összetett Poisson-eloszlásúnak nevezzük, ha $\eta = M_1 + ... + M_N$, ahol N Poisson eloszlású változó, M_i-k pedig azonos eloszlású változók, amelyek egymástól és N-től is függetlenek.

Legyenek M_i-k logaritmikus eloszlásúak, azaz a generátorfüggvényeik $G_M(z) = \frac{log(1-qz)}{log(1-q)}$. Ekkor $\eta = M_1 + ... + M_N$ generátorfüggvénye (N továbbra is Poisson):

$G_\eta(z) = E[E(z^\eta \mid N)] = E[(G_M(z))^N] = G_N(G_M(z))$ (függetlenség miatt).

Beírva az ismert generátorfüggvényeket:

$G_\eta(z) = exp\{\lambda(\frac{log(1-qz)}{log(1-q)} - 1)\} = exp\{\lambda\frac{log(1-qz)}{log(1-q)} - \lambda\frac{log(1-q)}{log(1-q)}\} = (\frac{1-qz}{1-q})^{\lambda/log(1-q)} =$

$= (1 - \frac{q}{1-q}(z-1))^{(\frac{\lambda}{log(1-q)})}$,

vagyis η eloszlása $NB(-\frac{\lambda}{log(1-q)}, q)$. Tehát a negatív binomiális eloszlás összetett Poissonként is megkapható, ha a másodlagos eloszlás logaritmikus. Ez a gyakorlatban is fennállhat: például egy baleset esetén több különböző kár is keletkezhet. Ezek számát közelíthetjük logaritmikus eloszlással, és ekkor az előbbi modellt kapjuk.
Érdekességképp megemlítem, hogyha a negatív binomiális eloszlás r paramétere végtelenhez tart, q paramétere pedig 0-hoz úgy, hogy a $\lambda = r\frac{q}{1-q}$ szorzat konstans, akkor a határeloszlás Poisson lesz. Ennek belátásához helyettesítsünk $\frac{q}{1-q} = \lambda/r$ -t a negatív binomiális eloszlás generátorfüggvényébe és alkalmazzuk kétszer a L'Hospital szabályt:

$$\lim_{r\to\infty}(1-\tfrac{\lambda(z-1)}{r})^{-r} = exp\{\lim_{r\to\infty} -rlog[1-\tfrac{\lambda(z-1)}{r}]\} =$$
$$= exp\{\lim_{r\to\infty} \frac{[1-\lambda(z-1)/r]^{-1}\lambda(z-1)/r^{-2}}{r^{-2}}\} =$$
$$= exp\{\lim_{r\to\infty} \tfrac{r\lambda(z-1)}{r-\lambda(z-1)}\} = exp\{\lim_{r\to\infty}[\lambda(z-1)]\} = exp\{[\lambda(z-1)]]\} \quad ,$$

ami a λ paraméterű Poisson eloszlás generátorfüggvénye.
Megjegyzésként, ha a negatív binomiális eloszlás esetén $r = 1$, akkor speciális esetként a geometriai eloszlást kapjuk. A geometriai eloszlás örökifjú tulajdonságát a következőképpen értelmezhetjük: tudván hogy legalább m darab kár már bekövetkezett, a további károk számának eloszlása nem függ m-től.

Tegyük fel, hogy van n darab szerződés, és mindegyik esetén p annak a valószínűsége, hogy bekövetkezik egy káreset (egymástól függetlenül). Ez akár előfordulhat életbiztosítás esetén, ahol az egyes biztosítottak azonos halálozási kategóriába tartoznak (például p annak a valószínűsége, hogy az adott személy a következő évben meghal). Ekkor az n szerződés esetén a káresetek száma binomiális eloszlást követ.
A binomiális eloszlás csak véges számú értéket vehet fel, ami szintén hasznos lehet, például egy autóbaleset esetén a megsérült személyek száma követhet ilyen eloszlást. Továbbá egyes esetekben a károk számára reális, hogy van egy felső korlát: gyakorlatilag elképzelhetetlen, hogy egy év alatt egy járművel több, mint 15 káreset történjen.
Az előző két eloszlással ellentétben a binomiális eloszlás szórásnégyzete kisebb, mint a várható értéke, ezért olyan adathalmazok esetén, melyeknél a tapasztalati átlag nagyobb a megfigyelt szórásnégyzetnél megfelelő lehet a binomiális eloszlás alkalmazása.

3.2. Az (a,b,0) osztály

N valószínűségi változó $(a, b, 0)$ eloszlású, ha $P(N = n) = (a + \frac{b}{n})P(N = n - 1)$, $n = 1, 2, ...$ (a $P(N = 0)$ valószínűséget a rekurzióból meg tudjuk kapni, ugyanis a valószínűségek összege 1 kell hogy legyen).
N pontosan akkor $(a, b, 0)$ eloszlású, ha Poisson, binomiális vagy negatív binomiális eloszlású. Ennek a bizonyítása megtalálható a [1]-ben, a következő táblázatban összefoglalom az eddigi

eloszlások esetén az a, b és $p_0 = P(N = 0)$ értékeket:

Eloszlás	a	b	p_0
Poisson	0	λ	$e^{-\lambda}$
Binomiális	$-\frac{p}{1-p}$	$(n+1)\frac{p}{1-p}$	$(1-p)^n$
Negatív binomiális	q	$(r-1)q$	$(1+\frac{q}{1-q})^{-r}$
Geometriai	q	0	$(1+\frac{q}{1-q})^{-1}$

A fenti rekurziót átírva $k\frac{p_k}{p_{k-1}} = ak + b$, $k = 1, 2, ...$, ahol $p_k = P(N = k)$.
Vagyis a baloldali kifejezés k-nak lineáris függvénye. A táblázat alapján, ha az egyenes a meredeksége 0, akkor Poisson eloszlás, ha pozitív, akkor negatív binomiális (ideértve a geometriait) és ha negatív, akkor binomiális eloszlás. Ez segíthet egy adott adatsor esetén megtalálni a megfelelő kárszám eloszlást. Ugyanis a $\frac{p_k}{p_{k-1}}$ hányadost közelíthetjük az $\frac{n_k}{n_{k-1}}$ hányadossal, ahol n_k azon szerződések száma, amelyek esetén k darab kár történt. És ha ezek az értékek közelítőleg egy egyenes mentén helyezkednek el, akkor célszerű lehet egy $(a, b, 0)$ eloszlással modellezni. Az egyenes meredekségének előjele pedig tovább szűkítheti a lehetséges eloszlások körét. Természetesen ez az eljárás nem végezhető el, ha bármely $n_k = 0$. Ebből kifolyólag ez pontatlan lehet és kevésbé hasznos akkor, ha a megfigyelések száma kicsi. Továbbá előfordulhat, hogy ugyan a meredekség pozitív vagy negatív, de nem tér el jelentősen 0-tól. Ezt megítélni nehéz, ilyenkor célszerű mindkét lehetséges eloszlást illeszteni, és valami jobban működő tesztet elvégezni annak érdekében, hogy eldöntsük melyik a megfelelőbb.

3.3. További kárszám eloszlások

Vannak olyan biztosítások, ahol ritkán következik be kár, azaz a legnagyobb annak a valószínűsége, hogy 0 kár lesz. Másrészt előfordulhat, hogy csak szigorúan pozitív számláló eloszlások jöhetnek szóba (például ha egy balesetnél akarja valaki a bekövetkezett károk számát modellezni). Ezért fontos lehet külön figyelmet fordítani p_0-ra, azaz arra a valószínűségre, hogy nem lesz káreset. Ezt az eddigi eloszlások nem feltétlen kezelik megfelelően, ezért apróbb módosításokat kell tenni.
N valószínűségi változó eloszlása (a, b) eloszlású, ha $P(N = n) = (a + \frac{b}{n})P(N = n - 1)$, $n = 2, 3, ...$
Az (a, b) eloszlás csupán annyiban tér el az $(a, b, 0)$ eloszlástól, hogy a rekurziót ebben az esetben csak $n = 2$-től követeljük meg. Mivel a valószínűségek összege 1 kell hogy legyen, ezért van lehetőség p_0 értékét megszabni ($0 \leq p_0 < 1$). Az (a, b) osztálynak értelemszerűen alosztálya az $(a, b, 0)$ osztály.

Egy másik fontos alosztálya a szigorúan pozitív eloszlások, más néven az $(a, b, 1)$ eloszlások:
N valószínűségi változó $(a, b, 1)$ eloszlású, ha (a, b) eloszlású és $P(N = 0) = 0$, azaz nem fordulhat elő az, hogy nem történik káreset.
Például a logaritmikus eloszlás az $(a, b, 1)$ osztályba tartozik, de nem tartozik az $(a, b, 0)$ eloszlások közé. M logaritmikus eloszlású, ha $P(M = k) = \frac{-1}{log(1-p)} \frac{p^k}{k}$, $k \geq 1$. Azaz $P(M = 0) = 0$ triviálisan teljesül. Továbbá

$\frac{P(M=k)}{P(M=k-1)} = \frac{p}{k}(k-1) = p - \frac{p}{k}$, ha $k \geq 2$.

Vagyis M (a, b) osztálybeli $a = p$ és $b = -p$ paraméterekkel (és így $(a, b, 1)$-beli is).

Korábban a negatív binomiális eloszlásnál volt szó arról, hogy előáll, mint összetett Poisson, mégpedig logaritmikus eloszlású másodlagos változókkal. A károk számának becslésére nemcsak az összetett Poisson, hanem egyéb összetett eloszlások is hasznosak lehetnek. Hiszen ha N egy számláló változó, amely generátorfüggvénye $G_N(z)$ és $M_1, M_2, ...$ független, azonos eloszlású valószínűségi változók $G_M(z)$ generátorfüggvénnyel, akkor a véletlen összeg $S = M_1 + ... + M_N$ generátorfüggvénye $G_S = G_N(G_M(z))$ (volt korábban bizonyítva). Az összetett eloszlások a biztosításban azért jelennek meg gyakran, mert N lehet a balesetek száma, M_i-k pedig az egyes baleseteknél előforduló károk száma (sérülések, autók, sérült személyek száma stb.). Megjegyzem, hogy mivel a balesetek számának eloszlása többnyire Poisson eloszlást követ, ezért az összetett eloszlások közül az összetett Poisson kiemelt fontosságú.
Legyen N és M Poisson eloszlású valószínűségi változó. Az összetett eloszlás generátorfüggvénye $G_S(z) = exp\{\lambda_1 exp\{\lambda_2(z-1)\}-1\}$, hiszen a Poisson eloszlás generátorfüggvénye $G_{Poi}(z) = exp\{\lambda(z-1)\}$. Ezt az új eloszlást A típusú Neyman eloszlásnak vagy Poisson-Poisson eloszlásnak nevezik.
Egy másik fontos összetett Poisson eloszlás, amikor a másodlagos változó binomiális eloszlású (Poisson-binomiális), aminek a generátorfüggvénye $G_S(z) = exp\{\lambda[1 + p(z-1)^m - 1]\}$. A binomiális eloszlás másodlagos eloszlásként megfelelő lehet például ha egy autóbalesetnél tekintjük a sérült személyek számát. Ugyanis feltehető, hogy közel ugyanakkora valószínűséggel sérülnek meg a jármű utasai.
Amikor az összetett Poisson eloszlásban a másodlagos változó logaritmikus, akkor negatív binomiális eloszlást kaptunk, ami nem egy új eloszlás az eddigiekhez képest, hiszen ez tagja az $(a, b, 0)$ osztálynak. Hasonlóan a geometriai-geometriai eloszlásról (melynek generátorfüggvénye $G_S(z) = [1 - \frac{q}{1-q}(\frac{1}{1-\frac{q}{1-q}(z-1)} - 1)]^{-1}$) is megmutatható, hogy (a, b)-beli, mégpedig egy geometriai eloszlással egyezik meg, melynek a 0-ban a valószínűsége módosított [3].
Az összetett Poisson eloszlás zárt a konvolúció alatt, azaz ha S_i-k független, összetett Poisson eloszlású valószínűségi változók λ_i Poisson-paraméterrel és $\{q_n(i) : n = 0, 1, 2, ...\}$ másodlagos változóval $(i = 1, 2, ...k)$, akkor az $S = S_1 + ... + S_k$ eloszlása szintén összetett Poisson, mégpedig a Poisson-paramétere $\lambda = \lambda_1 + ... + \lambda_k$ és a másodlagos változó $\{q_n : n = 0, 1, 2, ...\}$, ahol

$q_n = (\lambda_1 q_n(1) + ... + \lambda_k q_n(k))/\lambda$.

Bizonyítás: jelölje $Q_i(z) = \sum_{n=0}^{\infty} q_n(i) z^n \quad (i = 1, 2, ...k)$. Ekkor $G_{S_i} = E(z^{S_i}) = exp\{\lambda_i[Q_i(z) - 1]\}$. Mivel S_i-k függetlenek, ezért

$G_S(z) = \prod_{i=1}^{k} G_{S_i}(z) = \prod_{i=1}^{k} exp\{\lambda_i[Q_i(z) - 1]\} =$

$= exp\{[\sum_{i=1}^{k} \lambda_i Q_i(z) - \sum_{i=1}^{k} \lambda_i]\} = exp\{\lambda[Q(z) - 1]\}$,

ahol $\lambda = \sum_{i=1}^{k} \lambda_i$ és $Q(z) = \sum_{i=1}^{k} \lambda_i Q_i(z)/\lambda$, vagyis a generátofüggvény egyértelműsége miatt az állítást beláttuk.

Hasonlóan, mint a kárnagyságnál, a keverékeloszlások a kárszámok esetén is megjelennek, hiszen a populáció homogenitását továbbra is szükségszerű lehet figyelembe venni. Elsősorban a keverék Poisson eloszlások fontosak, ezért ezekre fektetek nagyobb hangsúlyt. Egy korábbi alfejezetben megmutattam, hogy a negatív binomiális eloszlás előáll mint keverék Poisson (gamma keverő változóval). Ha az M keverő változó szintén Poisson (μ paraméterrel), akkor a kapott eloszlás A típusú Neyman, ugyanis (hasonlóan a negatív binomiális eloszlásnál mutatott bizonyításhoz):

$G(z) = L_M(\lambda(1 - z)) = exp\{\mu(exp\{\lambda(z - 1)\} - 1)\}$, ami az A típusú Neyman generátorfüggvénye. Ez az eloszlás az összetett Poisson eloszlásoknál jött elő.

Az összetett és keverék Poisson eloszlások között egy fontos kapcsolat van. Egy eloszlás korlátlanul osztható, ha minden n-re előáll n darab független azonos eloszlás konvolúciójaként. Példa korlátlanul osztható eloszlásra a gamma, Poisson, negatív binomiális. A binomiális eloszlás nem korlátlanul osztható, mert csak véges különböző értéket vehet fel. Egy fontos tulajdonsága a keverék Poisson eloszlásoknak, hogy ha a keverő eloszlás korlátlanul osztható, akkor maga az eloszlás összetett Poisson is. Ezt a tételt nem bizonyítom, de példaként az előbb szerepelt az A típusú Neyman eloszlás és a negatív binomiális (keverék Poisson gamma keverő eloszlással, valamint összetett Poisson logaritmikus másodlagos eloszlászlással).

4. fejezet

Összkár modellezése

Eddig a kárszám és az egyes károk nagyságának eloszlásairól esett szó, és most az összkár modellezése a cél. Az összkár (S) egy adott időszak alatt történt károk nagyságának összege, azaz

$$S = X_1 + X_2 + ... + X_N \ ,$$

ahol N az időszak alatt megesett károk száma, X_i-k pedig az egyes károk nagyságai. Abban az esetben, ha $N = 0$, akkor S-et szintén definiáljuk 0-nak. X_i-kről általában feltehető, hogy független, azonos eloszlású változók (például ha S egy adott szerződés összkára, vagy egy adott kategóriába tartozó biztosítottak összkára). Gyakori feltevés, hogy N és az X_i kárnagyságok függetlenek. Vagyis N eloszlása nem függ az X_i kárnagyságoktól, továbbá feltételezve, hogy $N = n$, $X_1, ..., X_n$ eloszlása sem függ n-től. Ez sokat egyszerűsíthet az összkár megbecslésében (hiszen ekkor S egy összetett eloszlás), azonban nem biztos, hogy reális. Például könnyű elképzelni, hogy ha egy biztosított több káresetet okoz, akkor valószínűleg azok a károk kisebbek voltak. Később a szakdolgozat folyamán bemutatok egy modellt az összefüggő esetre is, egyelőre azonban a független esetet részletezem.

Ekkor S momentum generáló függvénye:

$$M_S(z) = E(e^{zS}) = E(exp\{z\sum_{i=1}^{N} X_i\}) =$$

$$= E(E(exp\{z\sum_{i=1}^{N} X_i\} \mid N))$$

$$= E(E(e^{zX_1}...e^{zX_N} \mid N)) = \qquad \text{(függetlenség itt használva)}$$

$$= E(\prod_{i=1}^{N} E(e^{zX_i} \mid N)) = E(E(e^{zX} \mid N)) =$$

$$= E(E(e^{zX})^N) = E(M_X(z)^N) =$$

$$= M_N(log(M_X(z))) \qquad \text{, (ahol } M_X(z) \text{ létezik).}$$

Mivel $E(S^n) = M_S^{(n)}(0) = \frac{d^n}{dz^n} M_S(z) \mid_{z=0}$, $(n \in \mathbb{N})$ ezért a S első két momentuma:

$$E(S) = \frac{d}{dz}\{M_N(log(M_X(z)))\} \mid_{z=0} =$$

$$= \{M_N'(log(M_X(z))) \times \frac{M_X'(z)}{M_X(z)}\} \mid_{z=0} =$$

$$= M'_N(log(M_X(0)) \times \frac{M'_X(0)}{M_X(0)} = M'_N(log(1)) \times \frac{M_X(0)}{1} =$$
$$= M'_N(0) \times M'_X(0) =$$
$$= E(N)E(X) \ ,$$

illetve

$$E(S^2) = \frac{d^2}{dz^2} M_S(z) \mid_{z=0} = \frac{d^2}{dz^2} \{M_N(log(M_X(z)))\} \mid_{z=0} =$$
$$= \frac{d}{dz} \{M'_N(log(M_X(z))) \times \frac{M'_X(z)}{M_X(z)}\} \mid_{z=0} =$$
$$= \{M''_N(log(M_X(z))) \times [\frac{M'_X(z)}{M_X}]^2 + M'_N(log(M_X(z))) \times \frac{M''_X(z)M_X(z) - [M'_X(z)]^2}{[M_X(z)]^2}\} \mid_{z=0} =$$
$$= M''_N(log(M_X(z))) \times [\frac{M'_X(0)}{M_X}]^2 + M'_N(log(M_X(0))) \times \frac{M''_X(0)M_X(0) - [M'_X(0)]^2}{[M_X(0)]^2} =$$
$$= M''_N(log(1)) \times [\frac{E(X)}{1}]^2 + M'_N(log(1)) \times \frac{E(X^2) \times 1 - [E(X)]^2}{1^2} =$$
$$= E(N^2) \times E(X)^2 + E(N) \times [E(X^2) - [E(X)]^2] =$$
$$= E(N^2)E(X)^2 + E(N)D^2(X) \ .$$

Vagyis S szórásnégyzete:

$$D^2(S) = E(S^2) - [E(S)]^2 = E(N^2)E(X)^2 + E(N)D^2(X) - [E(N)E(X)]^2 =$$
$$= E(X)^2[E(N^2) - [E(N)]^2] + E(N)D^2(X) =$$
$$= [E(Y)]^2 D^2(N) + E(N)D^2(X) \ .$$

Tehát az összkár eloszlásának várható értékét és szórását ki lehet fejezni a kárszám és a kárnagyság eloszlásának várható értékével és eloszlásával. Azokat pedig az előbbi két fejezetben tárgyaltuk, így ezek után a várható összkár becslésére ez a modell eléggé leegyszerűsödött. Mivel a kárszám eloszlása gyakran Poisson(λ) eloszlást követ (azaz S eloszlása összetett Poisson), ezért megjegyzem, hogy ekkor $E(S) = \lambda E(X)$ és $D^2(S) = \lambda E(X^2)$.

Az összkár modellezésének egy másik esete, amikor a szerződések száma fix (n), ekkor $S = X_1 + X_2 + ... + X_n$, ahol X_i-k az egyes szerződések esetén a károk nagysága (angolul individual risk model, [3]). Ez hasznos lehet, ha egy csoport (például n alkalmazott) biztosításáról van szó. A csoport tagjai különböző valószínűséggel okozhatnak kárt vagy más fedezeteik lehetnek, ezért X_i-kről nem szükséges feltenni, hogy független, azonos eloszlású változók. Ezzel a modellel a szakdolgozatban nem foglalkozok részletesebben, de fontosnak tartottam megemlíteni.

A következő példákban n darab szerződés van, mindegyikre megfigyelve a kárszám (N_i , $i = 1, 2, ..., n$) és az összkár (Y_i, $i = 1, 2, ..., n$).

1.) Tegyük fel, hogy az egyes káresetek egymástól (és a kárszámtól is) függetlenek, a kárnagyság exponenciális eloszlást követ μ paraméterrel, a károk száma pedig λ paraméterű Poisson eloszlást. A feladat λ és μ paraméterek becslése maximum-likelihood becsléssel.

A Poisson eloszlás λ paraméterét becsüljük a megfigyelt kárszámokból, a maximalizálandó likelihood függvény:

$$L(\lambda, \underline{N}) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{N_i}}{N_i!} \ ,$$

amiből a log-likelihood függvény:

$$l(\lambda, \underline{N}) = \sum_{i=1}^{n} -\lambda + N_i log(\lambda) - log(N_i!) \ .$$

Ezt szeretnénk maximalizálni, λ szerint deriválva és a kifejezést 0-val egyenlővé téve a következő egyenletet kapjuk:

$$\frac{dl}{d\lambda} = -n + \sum_{i=1}^{n} N_i/\lambda = 0 \qquad \Rightarrow \qquad \hat{\lambda} = \frac{\sum_{i=1}^{n} N_i}{n} \ .$$

Ami a μ paraméter becslését illeti, csak azokat a megfigyeléseket kell figyelembe venni, ahol történt kár. Az első fejezetben volt, hogy az exponenciális eloszlás a gamma eloszlás speciális esete ($Exp(\mu) \sim Gamma(1, 1/\mu)$, így az i-edik szerződés esetén az N_i darab exponenciális összege gamma eloszlást követ, amely paraméterei rendre $(N_i, 1/\mu)$ (feltéve, hogy a $N_i > 0$). Vagyis a likelihood és a log-likelihood függvény:

$$L\mu, \underline{Y}, \underline{N}) = \prod_{i:N_i>0} \frac{\mu^{N_i} Y_i^{N_i-1} e^{-\mu Y_i}}{(N_i-1)!} \ ,$$

$$l(\lambda, \mu, \underline{Y}, \underline{N}) = \sum_{i:N_i>0} N_i log(\mu) + (N_i - 1) log(Y_i) - \mu Y_i - log((N_i - 1)!) \ .$$

Ezt μ szerint deriválva és az egyenletet megoldva:

$$\frac{dl}{d\mu} = \sum_{i:N_i>0} N_i/\mu - \sum_{i:N_i>0} Y_i = 0 \qquad \Rightarrow \qquad \hat{\mu} = \frac{\sum_{i:N_i>0} N_i}{\sum_{i:N_i>0} Y_i}$$

2.) Az előző példában tegyük fel, hogy az i-edik szerződés t_i ideig áll fenn. Ekkor a kárszámok eloszlása Poisson marad, de λt_i paraméterrel (feltételezvén, hogy időben a károk egyenletesen jelennek meg). Megint becsüljük λ-t és μ-t!

Mivel a károk nagysága nem függ a megfigyelés hosszától, ezért a μ becslése ugyanúgy történik, mint az előző esetben, hiszen az összkár továbbra is gamma eloszlást követ. Vagyis

$$\hat{\mu} = \frac{\sum_{i:N_i>0} N_i}{\sum_{i:N_i>0} Y_i}$$

A λ paraméter becsléséhez a likelihood-függvény, majd a log-likelihood-függvény (itt most minden megfigyelést figyelembe kell venni, hiszen a kárszámot becsüljük):

$$L_{\underline{t}}(\lambda, \underline{N}) = \prod_{i=1}^{n} \frac{e^{-\lambda t_i} (\lambda t_i)^{N_i}}{N_i!}$$

$$l_{\underline{t}}(\lambda, \mu, \underline{Y}, \underline{N}) = \sum_{i=1}^{n} -\lambda t_i + N_i log(\lambda t_i) - log(N_i!) \ .$$

Ennek maximalizálásának érdekében most is a λ szerinti deriváltnak 0-val kell megegyeznie:

$$\frac{dl}{d\lambda} = \sum_{i=1}^{n} \frac{N_i}{\lambda t_i} t_i - \sum_{i=1}^{n} t_i = 0 \ .$$

Ebből

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} t_i} \ .$$

3.) Most az eddigieket módosítsuk annyival, hogy a kárszám $Poisson(\lambda t_i^a)$. Gyakorlati szempontból ez azért fontos, mert pontatlansághoz vezethet az a feltételezés, miszerint a káresetek időben egyenletesen történnek. Ezért pontosabb modellt eredményezhet egy plusz paraméter bevezetése a személyében.

Most a, λ és μ paraméterek becslése a feladat. Az eljárás ugyanaz, mint eddig, egyből a log-likelihood-függvény:

$$l_{\underline{t},a}(\lambda, \underline{N}) = \sum_{i=1}^{n} -\lambda t_i^a + N_i log(\lambda t_i^a) - log(N_i!) \ .$$

A λ és a szerinti deriváltak:

$$\frac{dl}{d\lambda} = \sum_{i=1}^{n} \frac{N_i}{\lambda} - \sum_{i=1}^{n} t_i^a = 0 \qquad \Rightarrow \qquad \lambda = \frac{\sum_{i=1}^{n} N_i}{\sum_{i=1}^{n} t_i^a}$$

$$\frac{dl}{da} = \sum_{i=1}^{n} -\lambda t_i^a log(t_i) + \sum_{i=1}^{n} N_i log(\lambda t_i) = 0 \qquad \Rightarrow \qquad \lambda \sum_{i=1}^{n} t_i^a log(t_i) = \sum_{i=1}^{n} N_i log(\lambda t_i) \ .$$

Két egyenletet kaptunk a két ismeretlenre (a és λ), az egyenlet (numerikus) megoldásával most nem foglalkozok.

Mivel csak a kárszám eloszlását változtattuk, a μ becslése továbbra sem változik,

így továbbra is $\hat{\mu} = \frac{\sum_{i:N_i<0} N_i}{\sum_{i:N_i<0} Y_i}$.

4.) Eddig a szerződések egyes káreseteinél a károk nagysága exponenciális volt. Most legyenek a kárnagyságok lognormális eloszlásúak (m és σ^2 paraméterekkel). Ellentétben az eddigivel, a lognormálisok összege nem hozható zárt alakra. Így a maximum-likelihood becslésben az összkár sűrűségfüggvényét nem tudjuk használni. A paraméterek becslésére tehát más módszer kell. Próbálkozzunk a momentum-módszer segítségével!

A λ Poisson paraméter becslése egyszerűen a megfigyelt kárszámok átlagából történik, azaz:

$$E(N_i) = \lambda = \frac{\sum_{i=1}^{n} N_i}{n} = \bar{N} \qquad \Rightarrow \qquad \hat{\lambda} = \bar{N} \ .$$

Az összkár egy összetett Poisson eloszlás, lognormális másodlagos változóval. Az S összetett (λ) Poisson eloszlás várható értékéről és szórásáról láthattuk feljebb, hogy ha a másodlagos változó X, akkor $E(S) = \lambda E(X)$ és $D^2(S) = \lambda E(X^2)$. Az első fejezetben pedig a lognormális eloszlás

momentumait kiszámoltam, amiből tudjuk, hogyha $logX \sim N(m, \sigma^2)$, akkor $EX = e^{m+\sigma^2/2}$ és $E(X^2) = e^{2m+2\sigma^2}$. Ezeket összerakva a momentum módszer két egyenlete:

$E(Y) = \lambda E(X) = \lambda e^{m+\sigma^2/2} = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}$,

$E(Y^2) = \lambda E(X^2) = \lambda e^{2m+2\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} = S^2$,

ahol S^2 a tapasztalati szórásnégyzet, λ pedig ismertnek tekintett a korábbi becslése által. Kis számolással ezt az egyenletrendszert megoldva m-re és σ^2-re a következőeket kapjuk:

$e^{\hat{\sigma}^2} = \bar{N}\frac{S^2}{\bar{Y}}$ és $e^{\hat{m}} = \frac{\bar{Y}^2}{\sqrt{\bar{N}^3 S^2}}$.

5.) Eddig voltak megfigyelések a kárszámokat és az összkárt illetően is. Most tételezzük fel, hogy csak az összkár (Y) van megfigyelve, az egyes károk eloszlása pedig kövessen exponenciális eloszlást μ paraméterrel. A károk száma (N) továbbra is Poisson(λ).

Hasonlóan a korábbi esetekhez, az összkár eloszlása ekkor is gamma, azonban az első paraméter most egy Poisson eloszlás eredménye, azaz az összkár egy keverék eloszlás. A keverék eloszlások momentumait kiszámoltam az első fejezetben, mégpedig ha X egy keverék eloszlás Λ keverő változóval, akkor $E(X^k) = E[E(X^k \mid \Lambda)]$. Esetünkben $Y \mid N$ gamma eloszlású $(N, 1/\mu)$ paraméterekkel és $N \sim Poi(\lambda)$. Így $E(Y \mid N) = N/\mu$, amiből

$E(Y) = E[E(Y \mid N)] = \lambda/\mu$,

valamint $E(Y^2 \mid N) = \frac{N(N+1)}{\mu^2}$, amiből

$E(Y^2) = E[E(Y^2 \mid N)] = \frac{1}{\mu^2}E(N^2 + N) = \frac{1}{\mu^2}((\lambda^2 + \lambda) + \lambda) = \frac{1}{\mu^2}(\lambda^2 + 2\lambda)$.

Így a momentum módszer egyenletrendszere a következő:

$E(Y) = \lambda/\mu = \frac{\sum_{i=1}^n Y_i}{n} = \bar{Y}$

$E(Y^2) = \frac{1}{\mu^2}(\lambda^2 + 2\lambda) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n} = S^2$.

Az egyenletrendszert megoldva μ-re és λ-ra, kapjuk:

$\hat{\lambda} = \frac{2\bar{Y}^2}{S^2 - \bar{Y}^2}$ és $\hat{\mu} = \frac{2\bar{Y}}{S^2 - \bar{Y}^2}$.

5. fejezet

Általánosított lineáris modell

Az általánosított lineáris modell (generalized linear model, GLM) célja, hogy egy megfigyelt véletlen változó várható értékét (és szórását) megjósoljuk, a megfigyelések bizonyos jellemzőinek hatását belevéve a modellbe. Az i-edik megfigyelés értéke legyen Y_i, a jellemzőit pedig pedig $X_{i,j}$-vel jelöljük. Ekkor Y egy vektor ("válasz változó") és X egy mátrix ("magyarázó változók"). A GLM a klasszikus lineáris modell általánosítása, ezért először röviden tekintsük át azt a [2] alapján.

5.1. Klasszikus lineáris modell

A klasszikus lineáris modellben feltesszük, hogy a megfigyelések függetlenek, és a várható érték a magyarázó változók lineáris kombinációjaként áll elő. Az alakja:

$$Y = X^T\beta + \varepsilon \ ,$$

ahol β a lineáris kombináció együtthatóiból álló vektor és ε egy hibatag, amelyről feltételezzük, hogy $N(0, \sigma^2 I)$ eloszlást követ. Más szóval $Y \sim N(X^T\beta, \sigma^2 I)$. Vagyis X csak a várható értéket befolyásolja, Y egyes komponensei esetén a szórásnégyzet megegyezik.
Összefoglalva a klasszikus lineáris modell feltevéseit:

- Y egyes komponensei függetlenek, normális eloszlásúak μ várható értékkel. A komponensek szórásnégyzetei megegyeznek és konstansok.
- A kapcsolat a várható érték és a magyarázó változók között: $\mu = X^T\beta$.

Az $\eta = X^T\beta$ lineáris kombinációt lineáris prediktornak nevezik. Ha a várható értéket a lineáris prediktor függvényeként akarjuk felírni, akkor ebben az esetben ez a függvény az identitás. Ez a megjegyzés most feleslegesnek tűnhet, de a GLM-ben ezt majd általánosítjuk.
Azonban ezeket a feltételeket nehéz megkövetelni a gyakorlatban. A normális eloszlás és a konstans szórásnégyzet túl szigorú feltétel lehet. Például nem megszokott, hogy a kárnagyság

eloszlása normális eloszlást kövessen, ha pedig a kárszámot szeretnénk modellezni, akkor már az is problémát okoz, hogy annak eloszlása diszkrét. Másrészt előfordulhat, hogy a válasz változó például csak szigorúan pozitív értékeket vehet fel, amit sért a normális eloszlás feltevése. Továbbá a második feltétel egyfajta additivitást eredményez a jellemzők hatásai közt, ami szintén problémás lehet: ha egy pillangó szárnyfelületét szeretnénk becsülni és a két magyarázó változó a szárny szélessége és hossza, akkor az additivitás helyett inkább a két változó szorzata lehet lényeges. Fontosabb, hogy a válasz több biztosítási alkalmazás esetén is inkább multiplikatívan változik a magyarázó változókkal.

5.2. Általánosított lineáris modell

Az általánosított lineáris modellben ezeket a feltevéseket gyengítjük, ezáltal sokkal rugalmasabb, használhatóbb modellt kapva. Először is, ahelyett, hogy a várható érték a magyarázó változók egy lineáris kombinációja, a GLM-ben a lineáris prediktor lehet a várható érték egy nemlineáris függvénye is. Másrészt a válasz változóra nincs normalitási megkötés, hanem az exponenciális eloszláscsalád (EDF) tagja kell legyen. Az EDF-ről később részletesebben szó lesz. Továbbá a szórásnégyzet nem kell konstans legyen, hanem a szórásnégyzet változhat a várható érték függvényében (ez a tulajdonság az EDF-ből örökletes). Vagyis a várható érték modellezése egy GLM-mel magával vonja a szórásnégyzet modellezését is.
Analóg módon a klasszikus lineáris modellekhez, összefoglalva a feltételezések a GLM esetén [2]:

- Y egyes komponensei függetlenek, eloszlásuk valamilyen EDF-beli eloszlás.
- A kapcsolat a várható érték és a magyarázó változók között:
  $$g(E(Y)) = g(\mu) = \eta = X^T\beta \ ,$$
  ahol g monoton, folytonosan differenciálható függvény és $Dom(g) = \mathbb{R}$.

Mivel alapvetően a $\mu(X)$ várható érték érdekel bennünket, ezért a második pontban szereplő kifejezést szokás $E(Y) = \mu = g^{-1}(\eta) = g^{-1}(X^T\beta)$ alakban írni (mivel g monoton és folytonosan differenciálható, ezért létezik g^{-1}). Tehát a cél nem más, mint β paraméterek becslése.
A normális eloszlás tagja az exponenciális eloszláscsaládnak és ha g az identitás, akkor pont a klasszikus lineáris modellt kapjuk. Tehát az valóban a GLM speciális esete.
A GLM tehát megengedi, hogy nem-normális megfigyeléseket is modellezzünk, ami különösen hasznos a biztosításban. Például a korábban szerepelt eloszlások a kárnagyságra és a kárszámra nem normális eloszlást követtek, de többnyire az EDF tagjai voltak. Vagyis külön-külön modellezhetjük a kárnagyságot és a kárszámot, amiből pedig az összkárt tudjuk becsülni (ha a kárszámok és a kárnagyságok függetlenek).

5.2.1. Kapcsolati függvény

A klasszikus lineáris modellben g az identitás volt, azaz a μ várható érték $\eta = X^T\beta$ -val. $X^T\beta$ egy lineáris kombináció, amely értéke bármely valós számot felveheti. Vagyis μ is bárhova eshet a valós számegyenesen. A gyakorlatban ez problémákhoz vezethet: ha például a válasz változónk szigorúan pozitív, akkor nem szeretnénk hogy a várható értéke negatív legyen. Például ha kárnagyságról van szó és gamma eloszlást feltételezünk, akkor a várható értéke a $(0, \infty)$ intervallumba kellene essen, hiszen a gamma eloszlás sűrűségfüggvénye csak itt pozitív. Másik szemléletes példa, ha Bernoulli (indikátor) eloszlást feltételezünk a válasz változónak, ekkor ugyanis a várható érték a $(0, 1)$ intervallumba kell essen. A GLM esetén ez a probléma kiküszöbölhető, mégpedig a kapcsolati függvény segítségével. Ugyanis a várható érték csak g^{-1} értékkészletébe eshet. Vagyis g-t megfelelően kell megválasztanunk annak érdekében, hogy használható modellt kapjunk (nem csak egy ilyen g lehet).
Például ha Poisson vagy gamma eloszlást feltételezünk, akkor mivel az csak pozitív értékeket vehet fel, ezért természetes választás lehet a $g = log$ választás, azaz a logaritmikus link. Ha pedig az eloszlás Bernoulli, akkor g a $(0, 1)$ intervallumot kell képezze $\mathbb{R}$-re. Jó választás lehet $log(\frac{\mu}{1-\mu}) = X^T\beta$ (logit link) vagy $\Phi^{-1}(\mu) = X^T\beta$, ahol Φ a standard normális eloszlás eloszlásfüggvénye (probit link). Ez a példa azt mutatja, hogy többféle g is jó lehet. Az, hogy melyiket használjuk ízlés kérdése, lényegében nincs nagy különbség, hogy melyiket választjuk.

5.2.2. Exponenciális eloszláscsalád

Az exponenciális eloszláscsalád egy 2-paraméteres eloszláscsalád, amely magába foglalja a legtöbb eddigi eloszlást, amiről szó esett (Poisson, binomiális, exponenciális, gamma, stb.). Formálisan egy Y változónak az eloszlása EDF-beli θ kanonikus és $\phi > 0$ diszperziós paraméterrel, ha a sűrűségfüggvénye:

$$f_Y(y; \theta, \phi) = exp\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\} \ ,$$

ahol $b(.)$ és $c(. \ , .)$ ismert függvények. A $c(y, \theta)$ egy normalizáló tag, vagyis abban van szerepe, hogy a sűrűségfüggvény integrálja 1 legyen. Tehát az eloszlást lényegében $b(\theta)$ határozza meg. (Megjegyzem, hogy az [2] és [5]-ben az exponenciális család definíciója egy kicsit általánosabb: az exponenciális függvényen belüli tört nevezőjében ϕ-nek egy $a(\phi)$ függvénye szerepel).
Példaként megmutatom, hogy a normális és Poisson eloszlás benne van ebben az eloszláscsaládban.
A normális eloszlás $\theta = (\mu, \sigma^2)$ sűrűségfüggvénye:

$$f_\theta(y) = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{(y-\mu)^2}{2\sigma^2}] = exp[\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2))],$$

innen pedig leolvasva: $\theta = \mu$, $\phi = \sigma^2$, $b(\theta) = \frac{1}{2}\theta^2$.
A Poisson eloszlás esetében $f(y) = \frac{\lambda^y e^{-\lambda}}{y!} \qquad \lambda > 0, \quad y = 0, 1, 2, ...$
Ekkor $log(f(y)) = ylog(\lambda) - \lambda - log(y!)$, vagyis $f(y) = exp[ylog(\lambda) - \lambda - log(y!)]$, ahonnan már látszik, hogy $\theta = log(\lambda)$ paraméterezéssel $b(\theta) = e^\theta$, $\phi = 1$ és $c(y, \theta) = -log(y!)$.

A következőekben megmutatom, hogy ha $Y \sim EDF(\theta, \phi)$, akkor $E(Y)$ csak $b(.)$-től függ, mégpedig $E(Y) = b'(\theta)$, valamint a szórásnégyzet $D^2(Y) = \phi b''(\theta)$.
Bizonyítás ([5] alapján, kis módosítással): legyen $l(\theta, \phi; y) = log(f_Y(y; \theta, \phi))$ a log-likelihood függvény. Bizonyos regularitási feltételek mellett:

$$E[\frac{\partial}{\partial\theta} l(\theta, \phi; Y)] = E[\frac{\partial}{\partial\theta} log(f_Y(Y; \theta, \phi))] =$$
$$= \int \frac{\frac{\partial}{\partial\theta} f_Y(y;\theta,\phi)}{f_Y(y;\theta,\phi)} f_Y(y; \theta, \phi)dy =$$
$$= \frac{\partial}{\partial\theta} \int f_Y(y; \theta, \phi)dy =$$
$$= \frac{\partial}{\partial\theta} 1 = 0 \ .$$

Másrészt (Fischer-információ)

$$E[(\frac{\partial}{\partial\theta} log f_Y(Y; \theta, \phi))^2] + E[\frac{\partial^2}{\partial\theta^2} log f_Y(Y; \theta, \phi)] = 0 \ ,$$

ugyanis:

$$\frac{\partial^2}{\partial\theta^2} log f_Y(Y; \theta, \phi) = \frac{\partial}{\partial\theta}(\frac{\frac{\partial}{\partial\theta} f_Y(y;\theta,\phi)}{f_Y(y;\theta,\phi)}) = \qquad \text{(hányados deriváltjából)}$$
$$= \frac{\frac{\partial^2}{\partial\theta^2} log f_Y(Y;\theta,\phi)}{f_Y(Y;\theta,\phi)} - (\frac{\frac{\partial}{\partial\theta} f_Y(y;\theta,\phi)}{f_Y(Y;\theta,\phi)})^2 \ .$$

Ebből várható értéket véve $E[\frac{\partial^2}{\partial\theta^2} log f_Y(Y; \theta, \phi)] = -E[(\frac{\frac{\partial}{\partial\theta} f_Y(y;\theta,\phi)}{f_Y(Y;\theta,\phi)})^2] = -E[(\frac{\partial}{\partial\theta} log f_Y(Y; \theta, \phi))^2]$,
ugyanis:

$$E[\frac{\frac{\partial^2}{\partial\theta^2} log f_Y(Y;\theta,\phi)}{f_Y(Y;\theta,\phi)}] = \frac{\partial^2}{\partial\theta^2} \int 1dx = 0 \ .$$

Összegezve és egyszerűbb alakba írva az eddigiek:

$$E(\frac{\partial l}{\partial\theta}) = 0 \ ,$$
$$E[\frac{\partial^2 l}{\partial\theta}^2] + E[(\frac{\partial l}{\partial\theta}^2)] = 0 \ .$$

Itt most $l(\theta, \phi; y) = \frac{Y\theta - b(\theta)}{\phi} + c(Y; \phi)$, amiből $\frac{\partial l}{\partial\theta} = \frac{Y - b'(\theta)}{\phi}$.
Tehát $0 = E(\frac{\partial l}{\partial\theta}) = \frac{E(Y) - b'(\theta)}{\phi}$, amiből $\underline{E(Y) = \mu = b'(\theta)}$.
Továbbá $\frac{\partial^2 l}{\partial\theta^2} + (\frac{\partial l}{\partial\theta})^2 = -\frac{b''''(\theta)}{\phi} + (\frac{Y - b'(\theta)}{\phi})^2$, így az előzőekből:

$$\frac{Y - b'(\theta)}{\phi} = \frac{Y - E(Y)}{\phi} \Rightarrow 0 = -\frac{b''(\theta)}{\phi} + \frac{D^2(Y)}{\phi^2} \Rightarrow D^2(Y) = \phi b''(\theta) \ .$$

Ezzel a bizonyításnak vége.
Fontosnak tartom megjegyezni, hogy a szórásnégyzet a kanonikus paraméter θ függvénye, ezért a várható érték függvénye is. Pontosítva $D^2(Y) = \phi b''(\theta)$ és $\theta = b'^{-1}(\mu)$, azaz

$$D^2(Y) = \phi b''(b'^{-1}(\mu)) \equiv v(\mu) \ .$$

Vagyis amikor a várható értéket modellezem a GLM-mel, akkor egyúttal a szórásnégyzetet is.

5.2.3. Offset tag

Előfordulhat, hogy egyes magyarázó változók hatása ismert, és ezért ahelyett, hogy β paramétereket ezen változó függvényében becsülnénk, valahogy máshogy adjuk hozzá a modellhez a hatását. Ennek egy módja az úgynevezett offset tag ξ bevezetése az alábbi módon:

$$\eta = X^T\beta + \xi \ ,$$

vagyis a lineáris prediktor definícióját változtattuk meg. Ebből következően:

$E(Y \mid X) = \mu(X) = g^{-1}(\eta) = g^{-1}(X^T\beta)$.

Ez különösen fontos lehet, ha például a kárszám eloszlása Poisson, ahol az egyes esetekben a megfigyelés időtartama különböző lehet. Nyilván fontos különbség van aközött, hogy például járműbiztosítás esetén a vezető 3 kárt okozott 5 év alatt, vagy 25 év alatt. Ezt úgy vehetjük bele a modellbe, hogy hozzávesszük a megfigyelt időtartam logaritmusát a lineáris prediktorhoz, azaz minden megfigyelés esetén:

$E(Y \mid X) = g^{-1}(\sum_j X_{ij}\beta_j + \xi_i) = exp[\sum_j X_{ij}\beta_j + log(d_i)] = d_i exp[\sum_j X_{ij}\beta_j]$,

ahol d_i az i-edik megfigyelés ideje.

A GLM működését a következő konkrét példán fogom bemutatni (hasonló példa a [2]-ben található). Tekintsünk egy járműbiztosítási példát, ahol négy megfigyelésünk van. A vezetőket két szempont szerint két-két osztályba osztjuk, például nem (nő és férfi) és lakhely (városi és vidéki) szerint. Ekkor négy magyarázó változó van: X_1 férfi, X_2 nő, X_3 városi és X_4 vidéki. Ezek az X_i változók indikátorok, 0-1 értékeket vehetnek fel. A megfigyelt átlagos kárnagyságot a következő táblázat foglalja össze:

	Városi	Vidéki
Férfi	900	600
Nő	400	300

A válasz változó Y, az átlagos kárnagyság.

Mivel az egyes magyarázó változók összefüggenek, ezért az egyik magyarázó változó (például X_4) elhagyható (általában célszerű a leggyakoribbat elhagyni, de itt most ez nem számít). Így gondolhatunk a modellre úgy, hogy van egy átlagos válasz a férfiakra (β_1), egy átlagos válasz a nőkre (β_2), és nemtől függetlenül egy plusz hatása annak, ha a vezető városi (β_3).

Vagyis $Y = \begin{pmatrix} 900 \\ 600 \\ 400 \\ 300 \end{pmatrix}$, $\quad X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, $\quad$ és $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$.

Ezután a válasz változó eloszlását választjuk meg. Mivel kárnagyságról van szó, ezért válasszuk a gamma eloszlást (α és θ pataméterekkel). A gamma eloszláshoz szokás inverz kapcsolati függvényt választani, azaz

$$E(Y) = g^{-1}(X^T\beta) = \begin{pmatrix} g^{-1}(\beta_1+\beta_3) \\ g^{-1}(\beta_1) \\ g^{-1}(\beta_2+\beta_3) \\ g^{-1}(\beta_2) \end{pmatrix} = \begin{pmatrix} (\beta_1+\beta_3)^{-1} \\ (\beta_1)^{-1} \\ (\beta_2+\beta_3)^{-1} \\ (\beta_2)^{-1} \end{pmatrix} .$$

A gamma eloszlás log-likelihood függvénye:

$l(y;\alpha,\theta) = \sum_{i=1}^4 \alpha log(\frac{y}{\theta}) - \frac{y}{\theta} - log(y) - log(\Gamma(\alpha))$.

A gamma eloszlás várható értéke $\theta\alpha = \frac{1}{\sum X_{ij}\beta_j}$ az inverz kapcsolati függvény miatt, amiből $\frac{1}{\theta} = \alpha \sum X_{ij}\beta_j$. Ezt behelyettesítve, leosztva α-val majd elhagyva néhány konstanst:

$\tilde{l}(y,\theta) = \sum_{i=1}^4 y_i log(\sum X_{ij}\beta_j) - y_i \sum X_{ij}\beta_j =$

$$= log(900(\beta_1 + \beta_3)) - 900(\beta_1 + \beta_3) +$$
$$+ log(600\beta_1) - 600 * \beta_1 +$$
$$+ log(400(\beta_2 + \beta_3)) - 400(\beta_2 + \beta_3) +$$
$$+ log(300\beta_2) - 300 * \beta_2 \ .$$

Ahhoz, hogy ezt maximalizáljuk vegyük a β_i szerinti parciális deriváltakat és tegyük azokat egyenlővé 0-val:

$$\frac{\partial \tilde{l}}{\partial \beta_1} = 0 \Rightarrow \frac{1}{\beta_1+\beta_3} + \frac{1}{\beta_1} = 1500$$
$$\frac{\partial \tilde{l}}{\partial \beta_1} = 0 \Rightarrow \frac{1}{\beta_2+\beta_3} + \frac{1}{\beta_2} = 700$$
$$\frac{\partial \tilde{l}}{\partial \beta_1} = 0 \Rightarrow \frac{1}{\beta_1+\beta_3} + \frac{1}{\beta_2+\beta_3} = 1300 \ ,$$

amely egyenletrendszert megoldva a következő megoldást kapjuk β-ra:

$$\beta_1 = 0.001704, \qquad \beta_2 = 0.003194, \qquad \beta_3 = -0.000609 \ .$$

Nincs más hátra, mint visszahelyettesíteni β_i-ket. A következők a kapott értékek:

	Városi	Vidéki
Férfi	913.125	586.875
Nő	386.875	313.125

6. fejezet

GLM az összkár modellezésére

A következő fejezet célja, hogy az összkár becslését az általánosított lineáris modell keretein belül végezzük el. Annak érdekében, hogy egy személy biztosítása megfelelően legyen beárazva, a várható veszteségének becslését el kell végezni. Erre néhány jellemzőjét figyelembe véve lehet következtetni a korábbi adatok alapján, így alkalmas lehet a GLM-mel való modellezés. A fejezet alapjául a [4] és [5] szolgáltak.

A harmadik fejezetben esett szó arról, hogy ha a károk számát és a károk nagyságát függetlennek tekintjük, akkor az összkár modellezése lényegesen leegyszerűsödik. Ezért először ezzel az esettel foglalkozunk.

6.1. Független modell

Az i-edik szerződésben az S_i összkár legyen

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

ahol N_i a kárszám, $Y_{i,j}$, $(j = 1...N_i)$ a károk nagyságai. $Y_{i,j} \stackrel{d}{=} Y_i$ független, azonos eloszlású változók, amelyek eloszlása nem függ N_i-től. Abban az esetben, ha $N_i = 0$, akkor S_i-t szintén definiáljuk 0-nak.

Ahogy korábban láttuk is, ekkor

$E(S_i) = E(N_i)E(Y_i)$ és $D^2(S_i) = E(N_i)D^2(Y_i) + D^2(N_i)E(Y_i)^2$,

vagyis a várható összkár nagysága egyszerűen a várható kárszám és kárnagyság szorzata.

A szerződéshez tartozó magyarázó változókat jelölje X_i (vektor). X_i mellett a kárszám várható értéke legyen ν_i, a kárnagyságé μ_i. A kárszámra és a kárnagyságra is GLM-et illesztve, rendre g_1 és g_2 kapcsolati függvényekkel:

$$g_1(\nu_i) = \eta_1 = X_i^T \underline{\alpha} \qquad \Leftrightarrow \qquad \nu_i = g_1^{-1}(X_i^T \underline{\alpha})$$
$$g_2(\mu_i) = \eta_2 = X_i^T \underline{\beta} \qquad \Leftrightarrow \qquad \mu_i = g_2^{-1}(X_i^T \underline{\beta}) \ ,$$

ahol $\underline{\alpha}$ és $\underline{\beta}$ a regressziós paraméterekből álló vektorok, amiket a GLM-ből kaptunk. Jegyezzük meg, hogy nem feltétlen releváns X_i minden komponense mindkét modellben, így megelőzően $\underline{\alpha}$

és $\underline{\beta}$ megfelelő komponensei 0-ra állíthatók.
Vagyis a független modell esetén az összkár

$$E(S_i) \;=\; \nu_i\mu_i \;=\; g_1^{-1}(X_i^T\underline{\alpha}) \times g_2^{-1}(X_i^T\underline{\beta}) \quad .$$

Kitüntetett esetben, amikor mindkét GLM-ben a kapcsolati függvény a logaritmus függvény, akkor

$$log(\nu_i) = X_i^T\underline{\alpha} \qquad \Leftrightarrow \qquad \nu_i = exp(X_i^T\underline{\alpha})$$
$$log(\mu_i) = X_i^T\underline{\beta} \qquad \Leftrightarrow \qquad \mu_i = exp(X_i^T\underline{\beta})$$

azaz

$$E(S_i) \;=\; exp(X_i^T\underline{\alpha} + X_i^T\underline{\beta}) \quad .$$

Több előnye van annak, hogy logaritmikus kapcsolati függvényt használunk. Egyrészt garantálja, hogy a kárszám és a kárnagyság várható értéke is pozitív legyen, másrészt minden egyes jellemző hatása egy új szorzótag hozzáadását jelenti (így elég egyszerű a struktúrája).

A GLM-ben szereplő eloszlások az exponenciális eloszláscsalád tagjai. Beláttam, hogy ilyen eloszlásokra a szórásnégyzet a várható érték függvénye (mégpedig $D^2(X) = \phi V(\mu)$, ha X EDF-beli ϕ és θ paraméterekkel, μ várható értékkel). Vagyis felhasználva ezt és a GLM-ből kapott várható értékeket, az összkár szórásnégyzete:

$$D^2(S_i) = E(N_i)D^2(Y_i) + D^2(N_i)E(Y_i)^2 =$$
$$= \nu_i\phi_{Y_i}V_{Y_i}(\mu_i) + \phi_{N_i}V_{N_i}(\nu_i)\mu_i^2 \; .$$

Tehát a várható érték GLM-mel való becslése magával vonja a szórásnégyzet becslését is.
Példaként tekintsük azt, amikor a kárszám eloszlása Poisson eloszlást követ, a kárnagyságé pedig gamma eloszlást. Ezek EDF-beli eloszlások, a paraméterezésük a következő: a Poisson eloszlás esetén a diszperziós paraméter 1, a kanonikus paraméter ν, a gamma eloszlás esetén pedig a diszperziós paraméter legyen ϕ és a kanonikus paraméter μ_i. Az EDF-beli eloszlások szórásnégyzetére korábban volt formula, ezt alkalmazva

$$D^2(Y_i) = \phi V_{Y_i}(\mu_i) = \phi\mu_i^2 \text{ és } D^2(N_i) = V_{N_i}(\nu) = \nu \; ,$$

ugyanis a gamma eloszlás esetén $V_{Y_i}(\mu) = \mu^2$ és V_{N_i} az identitás ([5], 19-21. oldal). Vagyis ezek alapján a következőt kapjuk:

$$D^2(S_i) = \phi\nu\mu_i^2 + \nu\mu_i^2 = (\phi+1)\nu\mu_i^2 \; .$$

Másik módja lehet az összkár modellezésének, ha nem a kárnagyságokra, hanem azok átlagára, $\bar{Y_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{i,j}$ -ra illesztünk GLM-et. Ugyanis ekkor

$$E[\bar{Y_i}] = E[\tfrac{1}{N_i}\textstyle\sum_{j=1}^{N_i} Y_{i,j}] = \mu_i = E[Y_i] \; ,$$

vagyis a kárnagyság és az átlagos kárnagyság modellezése ekvivalens.

6.2. Összefüggő modell

A kárszámok és kárnagyságok közti függetlenség feltételezése azonban pontatlan eredményekhez vezethet bizonyos esetekben. Például motorkerékpár biztosítás esetén azoknál a vezetőknél, akik több kárt okoztak, rendre kisebb károk jelennek meg, mint a kevesebb kárt okozóknál.

A cél, hogy a GLM segítségével azt az esetet is modellezzük, amikor az $Y_{i,j}$ kárnagyságok és N_i kárszámok közötti összefüggőség megengedett. Adott N_i esetén az $Y_{i,j}$, $(j = 1...N_i)$ kárnagyságokról továbbra is feltesszük, hogy azonos eloszlású, független változók, de az eloszlásuk most függ N_i-től. A [4]-t követve a modellben az összefüggőséget úgy fogjuk számításba venni, hogy feltételezzük, hogy a kárnagyságok átlaga függ N_i-től. Ezt úgy fogjuk elérni, hogy $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{i,j}$ átlagos kárnagyságra illesztünk GLM-et, amelyben az egyik magyarázó változó N_i lesz.

Az összkár ekkor ugyanis a következő alakba írható:

$$S_i = \bar{Y}_i N_i \ .$$

Az összkár várható értéke most már nem feltétlenül egyenlő a kárnagyság és kárszám várható értékének szorzatával, ugyanis $E(S_i) = E[N_i E(\bar{Y}_i \mid N_i)] \neq E(N_i)E(\bar{Y}_i)$.

Először N_i-re és $\bar{Y}_i$-re illesztünk GLM-et, hogy megtudjuk X hatását a várható értékeikre. A kárszámot hasonlóan becsüljük, mint a független esetben. A változás az átlagos kárnagyság esetében van. Mivel ekkor N_i is magyarázó változó, ezért

$$g(E[\bar{Y}_i]) = g(\mu_i) = X^T \underline{\beta} + N_i\theta \ ,$$

ahol X azon magyarázó változók, amelyek a független esetben is voltak, g a kapcsolati függvény, $\mu_i = E(\bar{Y}_i)$ és a regressziós paraméterek vektora $(\underline{\beta}, \theta)$. N_i együtthatóját, $\theta \in \mathbb{R}$ -t nevezzük a kárnagyság és kárszám közti összefüggőségi foknak. Jegyezzük meg, hogy az itt szereplő β együtthatók nem ugyanazok, mint a független modellben szereplő β-k, mert N_i plusz magyarázó változó jelenléte befolyásolja a többi regressziós paramétert.

Ha a GLM-ben logaritmikus kapcsolati függvényt alkalmazunk, akkor

$$log(E(\bar{Y}_i)) = log(\mu_i) = X^T\beta + \theta N_i \ ,$$

amiből

$$E(\bar{Y}_i) = exp\{X^T\beta + \theta N_i\} = \mu_0 e^{\theta N_i} \ ,$$

ahol μ_0 a független modellben a kárnagyság várható értékének becslése.

Ebből az összkár várható értéke:

$$\begin{aligned} &E(S_i) = E[N_i E(\bar{Y}_i \mid N_i)] = \\ &= E[N_i \mu_0 e^{\theta N_i}] = \mu_0 E[N_i e^{\theta N_i}] = \\ &= \mu_0 E[\tfrac{\partial}{\partial\theta} e^{\theta N_i}] = \mu_0 \tfrac{\partial}{\partial\theta} E[e^{\theta N_i}] = \\ &= \mu_0 \tfrac{\partial}{\partial\theta} M_{N_i}(\theta) \ , \end{aligned}$$

ahol M_{N_i} az N_i momentumgeneráló függvénye, és feltételezzük, hogy a harmadik sorban az in-

tegrál és a deriválás felcserélhető. Speciálisan, ha N_i Poisson eloszlású (λ paraméterrel), akkor ez a felcserélés megtehető, ugyanis:

$$E(\tfrac{\partial}{\partial\theta}e^{\theta N_i}) = E(Ne^{\theta N}) = \textstyle\sum_{k=0}^{\infty} ke^{\theta k}\frac{\lambda^k e^{-\lambda}}{k!} =$$
$$= \lambda e^{\theta}\textstyle\sum_{l=0}^{\infty} e^{\theta l}\frac{\lambda^l e^{-\lambda}}{l!} = \lambda e^{\theta}e^{-\lambda}\sum_{l=0}^{\infty}\frac{(e^{\theta}\lambda)^m}{l!} =$$
$$= \lambda e^{\theta}e^{-\lambda}e^{e^{\theta}\lambda} = \lambda e^{\lambda(e^{\theta}-1)}e^{\theta} = \tfrac{\partial}{\partial\theta}e^{\lambda(e^{\theta}-1)} = \tfrac{\partial}{\partial\theta}E(e^{\theta N_i}),$$

ugyanis a λ paraméterű Poisson eloszlás momentum generáló függvénye $e^{\lambda(e^{\theta}-1)}$.
Mivel a kárszám N_i ugyanúgy van modellezve, mint a független esetben, ezért

$$\tilde{g}(E(N_i)) = \tilde{g}(\nu) = X^T\beta \quad \Leftrightarrow \quad \nu = \tilde{g}^{-1}(X^T\beta)\ .$$

Ha azt feltételezzük, hogy a kárszám Poisson eloszlást követ és a GLM-ben logaritmikus kapcsolati függvényt alkalmazunk, akkor

$$\nu = e^{X^T\alpha} \quad \text{és} \quad M_{N_i}(z) = exp\{\nu(e^z-1)\}\ .$$

Vagyis

$$E(S_i) = \mu_0\tfrac{\partial}{\partial\theta}M_{N_i}(\theta) =$$
$$= \mu_0\tfrac{\partial}{\partial\theta}exp\{\nu(e^{\theta}-1)\} =$$
$$= \mu_0\nu exp\{\nu(e^{\theta}-1)+\theta\}\ .$$

Összehasonlítva a független esettel, az összkár várható értéke egy szorzótagban különbözik. Erre az $exp\{\nu(e^{\theta}-1)+\theta\}$ tagra úgy tekintünk mint egy korrekció az összefüggés miatt. Megjegyzem, hogy az előbbi $E(S_i)$-re kapott kifejezés semmit nem használ $\bar{Y}_i$ eloszlásából, az egyetlen megkötés az volt rá, hogy az exponenciális eloszláscsalád tagja legyen. Azonban az fontos, hogy az $\bar{Y}_i$-re illesztett GLM esetén a kapcsolati függvény a logaritmus volt.

További megjegyzés, hogy ha $\theta = 0$, akkor a modell ekvivalens a független modellel. Egyrészt a kárszám kezelése ugyanaz a két modellben. Másrészt ebben az esetben a regressziós paraméterek megegyeznek a független esetben kapottakkal, hiszen mivel $\theta = 0$, ezért az adatok ugyanazokkal a magyarázó változókkal lesznek modellezve. A maradék korrekciós tag pedig $\theta = 0$ esetén egyenlő 1-gyel, így

$$E_D(S_i) = exp(X^T\beta)exp(X^T\alpha) = \mu_0\nu = E_I(S_i)\ ,$$

ahol az alsó indexbe tett D és I az összefüggő és független esetre utal.

Ha az átlagos kárnagyságra illesztett GLM-ben a θ egy lényeges regressziós paraméter, akkor levonhatjuk azt a következtetést, hogy $\bar{Y}_i$-t jelentősen befolyásolja az N_i kárszám. Azonban a várható értékre tett hatása nem olyan egyértelmű, mert azáltal, hogy N_i-t hozzávettük a magyarázó változókhoz, a többi regressziós paraméter megváltozik. Azaz az egyes magyarázó változók hatása más lehet, ha N_i benne van a magyarázó változók közt vagy sem.

A független esettel ellentétben az összkár szórásnégyzetét most nem lehet olyan egyszerűen kifejezni a kárszám és kárnagyság momentumaival. A teljes szórásnégyzet tétele alapján:

$$D^2(S_i) = D^2(E(S_i \mid N_i)) + E[D^2(S_i \mid N_i)] =$$
$$D^2(E(\bar{Y}_i N_i \mid N_i)) + E[D^2(\bar{Y}_i N_i \mid N_i)] = D^2(N_i E(\bar{Y}_i \mid N_i)) + E[N_i^2 D^2(\bar{Y}_i \mid N_i)]\ .$$

A következőekben feltételezzük, hogy a kárszám eloszlása Poisson, a kárnagyságok pedig gamma eloszlásúak. Ha a kárnagyságok gamma eloszlásának EDF-beli paraméterezésében a kanonikus paraméter μ_1 és a diszperziós paraméter ϕ, akkor a feltételes (N_i) átlagos kárnagyság is gamma eloszlású lesz μ_i kanonikus és $\frac{\phi}{N_i}$ diszperziós paraméterrel [5]. Az EDF-beli eloszlások szórásnégyzetére korában kapott kifejezésből

$$D^2(\bar{Y}_i \mid N_i) = \frac{\phi}{N_i} V(\mu_i) = \frac{\phi}{N_i}\mu_i^2 \ ,$$

mert a gamma eloszlás esetén $V(\mu) = \mu^2$ ([5], 19-21. oldal). Ennek segítségével az összkár szórásnégyzetére kapott kifejezést tovább egyszerűsíthetjük:

$$D^2(S_i) = D^2(N_i\mu_i) + E[\phi N_i \mu_i^2] \ .$$

Most felhasználjuk a GLM-ből, hogy $\mu_i = exp\{X^T\beta + \theta N_i\}$, azaz :

$$\begin{aligned} D^2(S_i) &= D^2(N_i exp\{X^T\beta + \theta N_i\}) + E[\phi N_i exp\{2X^T\beta + 2\theta N_i\}] = \\ &= exp\{2X^T\beta\} D^2(N_i exp\{\theta N_i\}) + \phi exp\{2X^T\beta\} E[N_i exp\{2\theta N_i\}] = \\ &= \mu_0^2 D^2(N_i exp\{\theta N_i\}) + \phi\mu_0^2 E[N_i exp\{2\theta N_i\}] \ . \end{aligned}$$

A következőkben ezt a kifejezést szeretnénk tovább egyszerűsíteni (továbbra is a feltételezésünk szerint a kárszám Poisson, a kárnagyság pedig gamma eloszlást követ). Ehhez az alábbiakat fogjuk használni (M_{N_i} a momentumgeneráló függvénye lesz N_i-nek):

1.)
$$\begin{aligned} E[N_i exp\{\theta N_i\}] &= E[\tfrac{\partial}{\partial\theta} exp\{\theta N_i\}] = \tfrac{\partial}{\partial\theta} E[exp\{\theta N_i\}] = \\ &= \tfrac{\partial}{\partial\theta} M_{N_i}(\theta) = \tfrac{\partial}{\partial\theta} exp\{\nu(e^\theta - 1)\} = \\ &= exp\{\nu(e^\theta - 1)\}\nu e^\theta = \\ &= \nu exp\{\nu(e^\theta - 1) + \theta\} \ . \end{aligned}$$

2.)
$$\begin{aligned} E[N_i exp\{2\theta N_i\}] &= E[\tfrac{1}{2}\tfrac{\partial}{\partial\theta} exp\{2\theta N_i\}] = \\ &= \tfrac{1}{2}\tfrac{\partial}{\partial\theta} M_{N_i}(2\theta) = \nu exp\{\nu(e^{2\theta} - 1) + 2\theta\} \ . \end{aligned}$$

3.)
$$\begin{aligned} D^2(N_i exp\{\theta N_i\}) &= E[N_i^2 exp\{2\theta N_i\}] - [E[N_i exp\{\theta N_i\}]]^2 = \qquad \text{(az 1.) alapján)} \\ &= E[\tfrac{1}{4}\tfrac{\partial^2}{\partial\theta^2} exp\{2\theta N_i\}] - \nu^2 exp\{2\nu(e^\theta - 1) + 2\theta\} = \\ &= \tfrac{1}{4}\tfrac{\partial^2}{\partial\theta^2} M_{N_i}(2\theta) - \nu^2 exp\{2\nu(e^\theta - 1) + 2\theta\} = \\ &= \tfrac{1}{4}\tfrac{\partial^2}{\partial\theta^2} exp\{\nu(e^{2\theta} - 1)\} - \nu^2 exp\{2\nu(e^\theta - 1) + 2\theta\} = \\ &= \nu^2 exp\{\nu(e^{2\theta} - 1) + 4\theta\} + \nu exp\{\nu(e^{2\theta} - 1) + 2\theta\} - \nu^2 exp\{\nu(e^{2\theta} - 1) + 2\theta\} \ . \end{aligned}$$

A $D^2(S_i)$-re kapott legutóbbi kifejezésbe behelyettesítve 2.)-t és 3.)-t kapjuk:

$$D^2(S_i) = \nu\mu_0^2[\nu exp\{(e^{2\theta} - 1) + 4\theta\} + (\phi + 1) exp\{\nu(e^{2\theta} - 1) + 2\theta\} - \nu exp\{\nu(e^\theta - 1) + 2\theta\}] \ .$$

Ez a hosszú kifejezés $\theta = 0$ esetben lényegesen leegyszerűsödik:

$$D^2(S_i) = \nu\mu_0^2[\nu exp(0) + (\phi + 1) exp(0) - \nu exp(0)] = (\phi + 1)\nu\mu_0^2 \ ,$$

ami megegyezik a szórásnégyzetre adott kifejezéssel a független esetben. Azaz a $\theta = 0$ eset a független modellt adja vissza.

7. fejezet

Alkalmazás

Ebben a fejezetben valós biztosítási adatokon alkalmazom a korábbi fejezetekben ismertetett módszereket. Az adathalmaz egyéves járműbiztosításokból áll 2004 és 2005 között. A megjelenő károk mindegyike gépjárműbaleseti kár, tehát például a lopásból vagy rongálásból adódó károk nincsenek figyelembe véve. A független és összefüggő esetben adott modellt is alkalmazom, majd azokat összehasonlítom, hogy lássuk mennyiben változik a modell az összefüggő esetben.

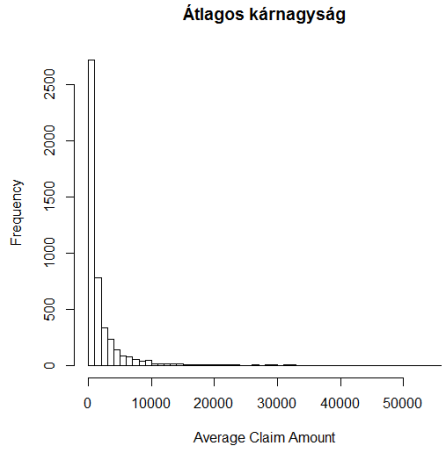
7.1. Adatok leírása

Az adathalmazban szereplő megfigyelések száma 67700, amelyek közül 4610 (6.81%) esetben történt legalább 1 kár. A kárnagyság becslése ez alapján 4610 megfigyelés alapján történik, hiszen csak a pozitív kárnagyságokat kell figyelembe venni (míg a kárszám becslésére a teljes adathalmazt használom). Mivel járműbiztosításról van szó, ezért nem meglepő az alacsony kárszám, legfeljebb 4 kár fordult elő a megfigyelt szerződések esetén. A kárszámok eloszlása (illetve függvényükben az átlagos kárnagyság) a következő táblázatban van összefoglalva:

Kárszám	Gyakoriság	Százalék	Átlagos kárnagyság
0	63090	93.19%	-
1	4321	6.38%	$ 1947.9
2	269	0.40%	$ 1477.1
3	18	0.03%	$ 1341.3
4	2	0.00%	$ 1109.7
Total	67700	100 %	$ 1917.7

Azt láthatjuk, hogy minél nagyobb a kárszám, annál kisebb az átlagos kárnagyság. Ez ma-

gyarázható például azzal, hogy ha valaki nagyobb balesetet szenved, akkor hajlamos kevesebb kárt okozni (akár nagyobb óvatosságból adódóan, vagy a jármű szervizelése miatt kevesebb ideig van megfigyelve). A GLM-et illetően ebből arra következtethetünk, hogy az összefüggő esetben a kárszám (mint magyarázó változó) regressziós együtthatója θ negatív. Ez amiatt van, hogy mivel kapcsolati függvényként a logaritmust alkalmazzuk, így az egyes magyarázó változók hatása egy plusz szorzóként jelenik meg e^{β} alakban (β a regressziós paraméter), ami pontosan akkor kisebb, mint 1, ha β negatív.
Az átlagos kárnagyság természetesen csak pozitív kárszám esetén van értelmezve. Legkisebb értéke 200$, míg a legnagyobb 55922.1$ volt. A következő ábra az átlagos kárnagyságok hisztogramja:



Mivel GLM-et szeretnénk alkalmazni, ezért az adatok közt szerepelnek minden egyes megfigyelés esetén a magyarázó változók is. Ezek esetünkben a jármű értéke(V_V), típusa (V_Body) és kora (V_Age), valamint a vezető neme (Gender), életkora (Age) és a lakhelye (Area). Előfordulhat, hogy egyes magyarázó változók között szorosabb kapcsolat van, például minél újabb egy autó, annál nagyobb az értéke. Ekkor nem akarjuk az adott magyarázó változók hatását többször belevenni a modellbe, mert az pontatlansághoz vezethet, így csak az egyiket hagyjuk meg. A következő táblázatba foglaltam a megfigyelt járművek értékét (V_V) és korát (V_Age). A jármű életkora kategorikus változóként szerepel az adatsorban (a kategóriák növekvően vannak: a legújabb járművek az 1-es kategóriában, a legidősebbek a 4-esben):

	V_V < 1	$1 <$V_V ≤ 1.5	$1.5 <$V_V ≤ 2	$2 <$V_V ≤ 3	$3 <$V_V	Összesen
V_Age 1	*3*	2175	2695	**3638**	**3705**	12216
V_Age 2	*156*	3942	**6084**	3078	3293	16556
V_Age 3	4769	**7301**	3023	3749	*1175*	20017
V_Age 4	**11769**	4223	2328	*539*	*55*	18914
Összesen	16697	17641	14130	11004	8228	67700

A táblázatból az látszik, hogy minél általában régebbi egy jármű, annál kisebb az értéke. Ezért szoros kapcsolatot feltételezvén a jármű kora és értéke között, a GLM illesztése során a jármű értékét kihagytam a magyarázó változók sorából.

7.2. Modellezés

A GLM illesztéséhez az R program $glm()$ beépített függvényét használtam. A kárszám esetén Poisson eloszlást, a kárnagyság esetén Gamma eloszlást feltételeztem. A kapcsolati függvény a logaritmus volt mind a kárszámnál, mind a kárnagyságnál. Számításba kell venni a megfigyelés idejét is az egyes szerződéseknél. Ennek érdekében egy offset tagot adtunk a modellhez a kárszám modellezése esetén. Ahogy korábban (a 4. fejezetben) is említettem, Poisson kárszám esetén a megfigyelési időt úgy tudjuk belevenni a modellbe, hogy a lineáris prediktorhoz hozzáadjuk az időtartam logaritmusát, azaz $E(Y) = exp\{X^T\beta + log(d_i)\} = d_i exp\{X^T\beta\}$, ahol d_i a megfigyelés időtartama. Az adatsorban a megfigyelési idő egy 0 és 1 közti szám, ami azt jelzi hogy az év mekkora hányadában volt megfigyelés alatt a szerződő. Vagyis az $exp\{X^T\beta\}$ tagot úgy is értelmezhetjük, mint a várható kárszám egy évre.
A kárnagyság modellezésekor azonban nem szükséges a megfigyelés idejét számításba venni, hiszen ott a káronkénti összeg számít, amire nincs hatása az időtartamnak.
Amikor a kárszámra illesztünk GLM-et, akkor minden megfigyelés súlya azonos (mindegyiké 1). (Megjegyzem, hogy a megfigyelés idejét súlyokkal is figyelembe lehetett volna venni offset tag helyett). Azonban a kárnagyságnál, amikor az átlagos kárnagyságot modellezzük, a károk számát (N_i) súlyként vesszük bele a modellbe.

Független modell

Ahogy korábban említettem, a károk számára Poisson eloszlás van feltételezve. A kárszámra először próbált modell[1]:

$$log(E(N)) = log(d_i) + \beta_1 VBody + \beta_2 VAge + \beta_3 Gender + \beta_4 Area + \beta_5 Age \ .$$

Erre GLM-et illesztve azt kapjuk, hogy a vezető nemének és lakhelyének hatása nem jelentős. Annak érdekében, hogy egy egyszerűbb modellt kapjunk, hagyjuk el ezeket a magyarázó változók sorából, és illesszünk így GLM-et az adatokra. Így már minden magyarázó változó lényegesnek bizonyul. Vagyis az egyszerűbb modell:

$$\nu_I = E(N) = d_i exp\{\beta_1 VBody + \beta_2 VAge + \beta_3 Age\} \ .$$

[1] A jármű típusa, kora és a vezető lakhelye, neme és kora kategóriákra van osztva az adatsorban, ezért ezek faktorokként jelennek meg itt.

A becsült regressziós paramétereket a következő táblázatba foglalom:

Regressziós paraméter	Becslés	Regressziós paraméter	Becslés
Intercept	-1.878	V_Body UTE	-0.190
V_Body COUPE	0.431	V_Age 1	0.089
V_Body HBACK	-0.061	V_Age 2	0.129
V_Body HDTOP	0.106	V_Age 4	-0.080
V_Body MCARA	0.581	Age 1	0.261
V_Body MIBUS	-0.039	Age 2	0.086
V_Body PANVN	0.066	Age 3	0.028
V_Body STNWG	0.038	Age 5	-0.220
V_Body TRUCK	-0.030	Age 6	-0.210

Megjegyzem, hogy mivel a megfigyelések jelentős részénél nem volt kár, ezért valószínűleg pontosabb lett volna, ha Poisson eloszlás helyett olyan diszkrét eloszlást feltételezünk, amely 0-ban módosított, azaz egy bizonyos p_0 valószínűséggel vesz fel 0-t, egyébként pedig mintha Poisson eloszlás lenne (csak az eloszlást módosítani kell p_0 miatt egy szorzóval). Ez az eloszlás egyébként tagja az (a,b,1) osztálynak.

A kárnagyságot illetően Gamma eloszlást feltételezünk. Ekkor a kárnagyságok átlaga is Gamma eloszlást követ. Először minden magyarázó változót használunk a GLM illesztésekor, vagyis

$$log(E[\bar{Y}]) = log(\mu_I) = \beta_1 VBody + \beta_2 VAge + \beta_3 Gender + \beta_4 Area + \beta_5 Age \ .$$

Lefuttatva a programot azt kapjuk, hogy a jármű típusán és korán kívül minden magyarázó változó hatása jelentős. Mindkét változót elhagyva a modell talán túlegyszerűsítetté válhat, ezért két új GLM-et illesztünk: egyszer a jármű típusát, majd az életkorát hagyjuk el a magyarázó változók sorából. A becsült regressziós paramétereket a két modellben a következő táblázat tartalmazza:

Regressziós paraméter	I,1	I,2
Intercept	7.478	7.42
V_Age 1	-0.086	
V_Age 2	-0.028	
V_Age 4	0.070	
Area A	-0.094	-0.093
Area B	-0.102	-0.104
Area D	-0.118	-0.108
Area E	0.081	0.087
Area F	0.287	0.312
Age 1	0.288	0.254
Age 2	0.093	0.087
Age 3	-0.008	-0.018
Age 5	-0.102	-0.100
Age 6	-0.039	-0.034
Gender Male	0.163	0.177
V_Body COUPE		0.322
V_Body HBACK		0.124
V_Body HDTOP		0.061
V_Body MCARA		-1.04
V_Body MIBUS		0.413
V_Body PANVN		0.158
V_Body STNWG		-0.003
V_Body TRUCK		0.191
V_Body UTE		0.067

Az összkárt illetően a független modellben

$$E(S_i) = E(\bar{Y}_i) \times E(N_i) = \mu_I \times \nu \ .$$

Összefüggő modell

Megint a kárszámra Poisson eloszlást, a kárnagyságra Gamma eloszlást feltételezünk.

A kárszám modellezése ugyanúgy történik, mint ahogy a a független esetben, tehát

$$\nu_D = \nu_I = E(N) = d_i exp\{\beta_1 VBody + \beta_2 VAge + \beta_3 Age\} \ .$$

Az összefüggő esetben az átlagos kárnagyság modellezésekor plusz egy magyarázó változó van az N_i kárszám által. Először megint az összes magyarázó változót használjuk:

$$log(E[\bar{Y}]) = log(\mu_{D,1}) = \beta_{1,1}VBody + \beta_{1,2}VAge + \beta_{1,3}Gender + \beta_{1,4}Area + \beta_{1,5}Age + \theta N$$

A félreértések elkerülése végett megjegyzem, hogy ezek a β paraméterek nem egyeznek meg a független esetbeli β paraméterekkel.

A GLM illesztésekor az derül ki, hogy a jármű típusának és korának a hatása gyakorlatilag megint elhanyagolható.. Ezért hasonlóan, mint a független kárnagyságnál, módosításokat végezve illesszünk még két modellt: először hagyjuk el a jármű típusát, majd a jármű korát a magyarázó változók sorából. Az átlagos kárnagyság becsült várható értékét jelölje $\mu_{D,1}$ és $\mu_{D,2}$ rendre. A becsült regressziós paramétereket megint egy táblázatban foglalom össze:

Regressziós paraméter	D,1	D,2
Intercept	7.717	7.641
N	-0.240	-0.232
V_Body COUPE		0.382
V_Body HBACK		0.133
V_Body HDTOP		0.061
V_Body MCARA		-1.05
V_Body MIBUS		0.396
V_Body PANVN		0.110
V_Body STNWG		0.12
V_Body TRUCK		0.196
V_Body UTE		0.093
Gender Male	0.165	0.178
V_Age 1	-0.095	
V_Age 2	-0.035	
V_Age 4	0. -0.240	
Area A	-0.082	-0.083
Area B	-0.091	0.096
Area D	-0.079	-0.071
Area E	0.082	0.086
Area F	0.283	0.306
Age 1	0.303	0.263
Age 2	0.100	0.083
Age 3	0.005	-0.004
Age 5	-0.104	-0.101
Age 6	-0.035	-0.028

Korábban említettük az adatok leírásánál, hogy azt várjuk, hogy minél több káreset történt, annál kisebb az átlagos kárnagyság. Ez teljesül is, ugyanis az N magyarázó változóhoz tartozó regressziós paraméter negatív, azaz $e^{\theta N} \leq 1$.

Mégegyszer, mivel a kapcsolati függvény a logaritmus, így $(j = 1, 2)$:

$$\mu_{D,j} = exp\{X^T \underline{\beta}_{D,j} + \theta N\} \ ,$$

azaz a módosított átlagos kárnagyság:

$$\mu^0_{D,j} = exp\{X^T \underline{\beta}_{D,j} = exp\{\beta_{j,0} + \beta_{j,1} VBody + \beta_{j,2} VAge + \\ + \beta_{j,3} Gender + \beta_{j,4} Area + \beta_{j,5} Age\} \ ,$$

ahol $\beta_{1,1} = 0$ és $\beta_{2,2} = 0$.

A korrekciós tag pedig:

$$exp\{\nu(e^\theta - 1) + \theta\} \ .$$

A j-edik összefüggő modell esetén az összkár várható értéke tehát a következő:

$$E[S_j] = \nu \mu^0_{D,j} exp\{\nu(e^\theta - 1) + \theta\} \ .$$

Modellek összehasonlítása

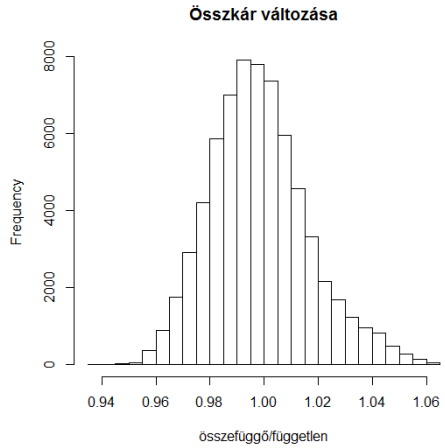
Ebben a részben a független és összefüggő modellt hasonlítom össze. A cél az összkár becslése. A kárszámra azt a modellt használom, amikor a vezető nemét és lakhelyét elhagyjuk a magyarázó változók sorából. Az átlagos kárnagyságot illetően pedig a független és összefüggő esetben is azt, amikor a jármű típusára vonatkozó változó van elhagyva.

A két modell közti különbség az, hogy átlagos kárnagyság modellezésében az N_i kárszámot is felhasználjuk magyarázó változóként. Ezáltal a többi magyarázó változóhoz tartozó regressziós paraméterek is megváltoznak. A legtöbb esetben a változás kicsi, a jármű életkorához tartozót kivéve a regressziós paraméterek jellemzően nőttek.

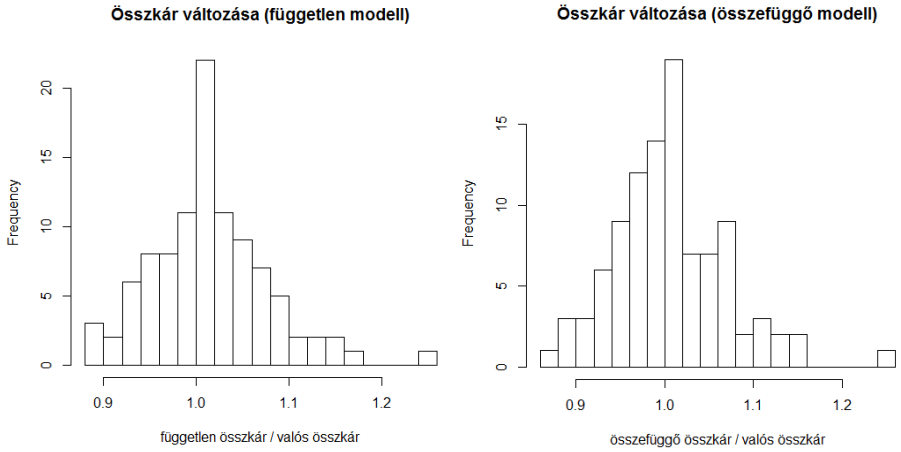
Például az intercept tag $\sim 3.2\%$ -kal nőtt (amikor összefüggő modellre váltottunk a független helyett). Ez látszólag nem tűnik jelentős változásnak, azonban nézzük az intercept tag különbségét a két modellben: $\beta_D - \beta_I = 0.239$. Mivel logaritmikus kapcsolati függvényt használunk az átlagos kárnagyság esetén is, ez $\sim +27\%$ -os növekedést jelent az átlagos kárnagyságban, ugyanis $e^{0.239} = 1,27$. Tehát a regressziós paraméterek megváltozása jelentős lehet, még akkor is, ha arányaiban nem változtak nagyon.

A független esetben az összkárt a kárszám és az átlagos kárnagyság becsült várható értékének a szorzataként kaptuk, amíg az összefüggő modellnél ez kiegészült egy korrekciós taggal. Áttérve az összefüggő modellre a függetlenről, átlagosan a becsült összkár 0.18% -kal csökkent a 47500 megfigyelt esetben, ami nem mondható jelentősnek. A legnagyobb növekedés 6.3% volt: UTE típusú, 2-es életkorú jármű, középkorú (agecat 5), férfi sofőr, B lakhely esetén. A legnagyobb csökkenés az összkár várható értékét illetően 6.2% volt: HDTOP típusú, 2-as életkorú jármű, fiatal (agecat 1), női sofőr, C lakhely esetén.

A becsült összkár változását az összefüggő esetre való áttéréskor a következő hisztogramon ábrázoltam:



Annak érdekében, hogy megvizsgáljuk, hogy pontosabb-e az összefüggő modell vagy sem, véletlenszerűen kiválasztjuk a megfigyeléseink 70%-át, azokra elvégezzük a GLM-et, majd a maradék 30% teljes összkárát becsüljük és megnézzük mennyiben tér el a valós adatoktól. Ezt 100-szor elvégezzük. Az átlagos eltérés a 100 futtatás után a független modellt alkalmazva 32605$, az összefüggő modell esetén pedig 8891$. A becsült teljes összkárok arányát a következő két hisztogramon ábrázolom:



Ha a becsült és valós összkárok átlagos négyzetes eltérését nézzük, abból jobban következtethetünk arra, hogy melyik lehet a pontosabb modell. Ezek gyökét tekintve a független modell alkalmazása esetén ez az érték 132300$, az összefüggő esetben pedig 124914$. Mivel ez az összefüggő modell esetén valamelyest kisebb, ezért mondhatjuk, hogy az összefüggőség figyelembevételével pontosabb modellt kaptunk.

A 100 esetben az összefüggő modell kivétel nélkül kisebb teljes összkárt jósolt, mint a független, átlagosan 0.87%-kal. A legnagyobb különbség a két becsült teljes összkár között legfeljebb 1,4% volt, ami 39598$-t jelentett. Ez viszonylag kis különbségnek mondható amellett, hogy a valós összkárt 293189$ (összefüggő) és 332778$ (független) hibával becsülték.

Tehát összességében mondhatjuk, hogy van különbség a két modell között és az összefüggőség figyelembevételével pontosabb modellt kapunk, de nem túl jelentős mértékben.

Összefoglalás

A biztosításban hagyományosan függetlennek tekintik a károk számát és nagyságát. Az összkár modellezését ez lényegesen leegyszerűsíti, ugyanis ekkor az összkár várható értéke megegyezik a kárnagyság és a kárszám várható értékének szorzatával. A szakdolgozat elején bemutattam a legjellemzőbb káreloszlásokat és tulajdonságaikat, valamint a kárszám modellezését is.

Azonban mivel a függetlenség ezek közt nem mindig tételezhető fel, szükség van egy modellre az összefüggő esetre is. Hasonlóan a független esethez, itt is külön-külön GLM-et illesztünk a kárszámra és az átlagos kárnagyságra. A kárszám modellezése ugyanúgy történik, mint a független esetben, azonban az átlagos kárnagyságnál a magyarázó változók sorához hozzávesszük a kárszámot is. Így az összkár várható értékére azt lehetett megmutatni, hogy a várható kárszám, egy módosított átlagos kárnagyság és egy a függetlenségért felelős korrekciós tag szorzataként áll elő, ha a kárszám eloszlása Poisson eloszlást követ és a GLM-ben logaritmikus kapcsolati függvényt alkalmazunk. Láttuk, hogy a független eset az összefüggőnek a speciális esete, hiszen olyankor a korrekciós tag pontosan 1, a módosított átlagos kárnagyság pedig megegyezik a független esetbeli átlagos kárnagysággal.

Egy valós adatsoron elvégezve a modellezést azt tapasztaljuk, hogy az összefüggő modellre áttérve a függetlenről a becsült összkár átlagosan csökkent, de csak kis mértékben. Továbbá a teljes adatsornak véletlenszerűen kiválasztott 70%-ára is GLM-et illesztettem és a maradék 30%-on becsültem a teljes összkárt, majd ezt még 100-szor elvégeztem. A becsült teljes összkárt összehasonlítva a valóssal azt lehet mondani, hogy az összefüggő modell valamivel pontosabb.

Mivel az összefüggő esetben az átlagos kárnagyságra illesztett GLM-ben a kárszám magyarázó változóhoz tartozó regressziós paraméter negatív, ezért az a feltételezésünk is beigazolódott, hogy a nagyobb károkat szenvedő vezetők rendre kevesebb kárt okoznak.

A használt adatsorban viszonylag kevés megfigyelésnél jelent meg kár. Ezért a Poisson eloszlás helyett talán pontosabb lett volna egy $(a, b, 1)$ osztálybeli eloszlást (például 0-ban módosított Poisson eloszlást) alkalmazni a kárszámra. Az összefüggő modellben azonban lényegesen kihasználtuk, hogy Poisson eloszlást feltételezünk (és hogy logaritmikus kapcsolati függvényt használunk), ezért további munkaként érdemes lenne a modellt bővíteni, hogy más eloszlások is használhatóak legyenek, vagy olyan adatsoron is elvégezni a modellezést ahol több esetben jelent meg kár.

Irodalomjegyzék

[1] Arató Miklós, *Nem-életbiztosítási matematika*, Eötvös kiadó, 2001.

[2] D. Anderson, S. Feldblum, C. Modlin, N. Thandic, *A Practitioner's Guide to Generalized Linear Models* , Watson Wyatt, 2007.

[3] Stuart A. Klugman, Harry H. Panjer, Gordon E. Willmot, *Loss Models: From Data to Decisions, II.*, Wiley, 2004.

[4] J.Garrido, C.Genest, J.Schulz, *Generalized linear models for dependent frequency and severity of insurance claims*, 2016, https://www.sciencedirect.com/science/article/pii/S0167668715303358

[5] Juliana Schulz, *Generalized Linear Models for a Dependent Aggregate Claims Model*, Concordia University Montreal, 2013, https://spectrum.library.concordia.ca/977691/1/Schulz_MSc_F2013.pdf

[6] G. Z. Heller, Jong P., *Generalized Linear Models for Insurance Data*, Cambridge University Press, 2008.

[7] Dan Ma, *A catalog of parametric severity models*, 2017, https://actuarialmodelingtopics.wordpress.com/tag/pareto-distribution/

Köszönetnyilvánítás

Ezúton szeretnék köszönetet mondani témavezetőmnek, Arató Miklósnak, hogy ismertette velem a témát, rengeteg hasznos tanáccsal és észrevétellel látott el, valamint bármikor alapos magyarázatot adott kérdéseimre.
Köszönöm családomnak és barátaimnak, akik a szakdolgozatom írása folyamán végig támogattak.