Eötvös Loránd Tudományegyetem

Természettudományi kar

Általánosított lineáris modellek a biztosításban

MSc Diplomamunka

Készítette:

Tóth András

Biztosítási és Pénzügyi

Matematika MSc

Aktuárius szakirány

Témavezető:

Szamoránsky János

Aegon Magyarország

Általános Biztosító Zrt.

Nem-életbiztosítás aktuárius

Budapest

2013

Tartalomjegyzék

Bevezetés

Diplomamunkám célja a biztosítási matematikában fontos szerepet betöltő általánosított lineáris modellek megismertetése a tisztelt Olvasóval. Hétköznapi példák segítségével a modell szisztematikus felépítése a megfelelő matematikai és statisztikai háttér bemutatásával, illetve a modell gyakorlatban történő alkalmazásával, amelyet a biztosítók elsősorban a biztosítási nettó díj meghatározására használnak.

Diplomamunkámat rövid történeti áttekintéssel kezdem a korábbi modellekről, azok hiányosságairól, felmerülő nehézségeikről illetve a vizsgálatokat megelőző feltételezésekről. Fokozatosan fogok eljutni az általánosított lineáris modellek szerkezeti felépítéséhez, a modell különböző részeihez illetve a modellhez kapcsolódó feltételezésekhez. A való életben előforduló példák segítségével fogom bemutatni a különböző modellek előnyeit illetve hátrányait, amelyhez a szabad forráskódú *R* programcsomagot hívom segítségül.

Emellett kitekintésként szó lesz az általánosított lineáris modellek más matematikai területekkel való kapcsolatáról többek között az exponenciális eloszlás-családról, kevert eloszlásokról közülük is elsősorban az összetett Poisson eloszlásról , a credibility elméletről és ezek különböző előnyös tulajdonságairól, amelyek elősegítik alkalmazhatóságukat a biztosítási matematika különböző területein.

Amikor az Olvasó már kellő ismeretanyaggal rendelkezik a modellről, áttérünk ennek a gyakorlati alkalmazására a biztosítási piacon. A modellt világszerte használják neméletbiztosítások nettó díjának meghatározásánál többek között a kötelező gépjármű felelősségbiztosítások árazásánál. A továbbiakban nettó díj alatt a kárszükségletet értem, ami egy szerződésre, egy év alatt várhatóan kifizetendő károk összege, költség, jutalék és

egyéb addicionális elem nélkül.

Itt szeretném megragadni az alkalmat, és szeretnék köszönetet mondani témavezetőmnek, Szamoránsky Jánosnak, aki segített megtalálni a diplomamunkám témájául szolgáló általánosított lineáris modellekkel kapcsolatos megfelelő színvonalú szakirodalmat, emellett idejét nem sajnálva a rendszeres konzultációk alkalmával rengeteg segítséget kaphattam tőle mind a modell matematikai hátterének megértésekor mind a modell biztosítási gyakorlatban történő alkalmazásakor, illetve az *R* programcsomag használatakor is hasznos tanácsokkal látott el.

1. fejezet

A modell kialakulása

Minden fejezetet szeretnék forrás megjelöléssel kezdeni, hogy egyértelművé tegyem az adott részek eredményeit és jelöléseit melyik irodalomjegyzékben található szakirodalom segítségével dolgoztam fel.
Az első fejezetben található különböző modellekkel kapcsolatos jelölésrendszer kialakításában az [1]-es szakirodalom nyújtott nagy segítséget, míg az itt megoldott számolási példák saját fiktív adatokon alapszanak.

1.1. Egyfaktoros analízis

A biztosítási matematikában használatos legtöbb fogalom általában angol nyelven szerepel a világirodalomban, így a legtöbbet nem is fordították le magyar nyelvre, vagy ha le is fordították kényelmesebb az eredeti angol elnevezést használni. Diplomamunkám készítése közben több ilyen szakkifejezéssel is találkoztam, ezeket a helyzettől függően próbáltam magyar nyelvre átültetni, amennyiben ez nem volt lehetséges, megtartottam az eredeti kifejezést. Ilyen például a *one-way analysis* is, amit *egyfaktoros analízis*–nek fordíthatnánk leginkább, de a szakirodalomban ezt az elnevezést nem igazán használják.
A módszer lényege, és egyben hiányossága is a következő. Legyenek adottak egy bizto-

sító adataiból, illetve tapasztalataiból képzett statisztikák (pl. az átlagkár vagy a kárgyakoriság). Az eljárás összegzi ezeket külön-külön minden egyes magyarázó változó szerint majd a létrehozott szegmensek kártapasztalatát vizsgálja tovább. A módszer minden esetben egyetlen változó hatását vizsgálja, amivel az a probléma, hogy a különböző változók közötti kapcsolatot, mint amilyen például a korreláció lehet, figyelmen kivül hagyja, ezzel pedig torzítja az eredményeinket.

1.1.1. A modell hiányossága

A könnyebb érthetőség végett, szeretném bemutatni a módszer hiányosságát egy példán keresztül. Kárigények alakulását vizsgáljuk a vezető, illetve a gépjármű kora alapján. Feltételezhetjük, hogy a fiatalabb gépjárművezetők általában - az anyagi hátterük miatt- idősebb (használt) gépjárművet vezetnek. A régóta használatban lévő járművek, illetve a fiatalabb, ezért kevesebb tapasztalattal rendelkező vezetők gyakrabban okoznak/szenvednek balesetet. A one-way analízis ezt a hatást kétszeresen mutatja ki, hiszen a régebbi gépjárműveket elsősorban fiatalabb sofőrök vezetik, így a vezető kora két különböző helyen is érezteti hatását. Mivel a gépjámű és a vezető kora között kapcsolat van (korreláció), ezeket a változók közötti rejtett összefüggéseket kezelnünk kell, hiszen ezeknek a tapasztalatoknak a felhasználása egy új termék díjkalkulációjához, vagy egy már meglévő termék utókalkulációjához, jelentős mértékben torzítani fogja az eredményeinket.

1.2. A klasszikus lineáris modell

A klasszikus lineáris modellt szinte csak egyetlen lépés választja el az általánosított lineáris modelltől, de ennek előzetes ismerete elengedhetetlen lesz a GLM[1] tárgyalásakor. A klasszikus lineáris modell célja, a korábbi modellek hiányosságainak kiküszöbölése, azaz a különböző magyarázó változók együttes hatásának vizsgálata a magyarázott válto-

[1]GLM - Generalized Linear Model, azaz általánosított lineáris modell. A továbbiakban ezt a rövidítést alkalmazom. Egyes szakirodalmakban még használatos a GLIM rövidítés is.

zóra vonatkozólag, az eredmények torzítása nélkül. A magyarázó változókat X-szel, míg a magyarázott változót Y-nal fogom jelölni. A klasszikus lineáris modell úgy tekint Y-ra mint egy determinisztikus változó és egy nulla várható értékű véletlen tag összegére.

$$Y = \mu + \varepsilon \tag{1.1}$$

Ahol μ *(a determinisztikus változó)* nem más, mint a magyarázó változók lineáris kombinációja. Jelöljük Y_i-vel a különböző realizációit a magyarázott Y változónak. Ekkor az i-dik realizáció, p számú magyarázó változó esetén, a következőképpen írható fel:

$$Y_i = \mu_i + \varepsilon_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i \tag{1.2}$$

A magyarázó változók egyes szintjeire, úgynevezett mérési pontként tekintünk, amik $0-1$ értéket vesznek fel attól függően, hogy az adott megfigyelés rendelkezik-e a változó által jelölt tulajdonsággal. Hogy mit tekintünk egy változó szintjének, azt kifejtem a modell gyakorlati alkalmazása során. Az ε pedig a mérésből eredő bizonytalanság, azaz a mérési hiba.

A modellel kapcsolatos feltevéseink a következők:

1. A magyarázó változók lineáris kombinációjaként írható fel μ_i, ami megfigyelésenként eltérhet

2. a hibatag ε normális eloszlású, 0 várható értékkel és σ^2 varianciával[2], azaz $\varepsilon \sim N(0, \sigma^2)$. A szórásnégyzet minden esetben azonos.

A modell alkalmazása során nem tudjuk minden megfigyelés esetében kizárni a mérési hibákat, viszont szeretnénk azokat minimalizálni. Az $i.$ mérési hibát a következőképpen fejezhetjük ki az (1.2)-es egyenletből:

$$\varepsilon_i = Y_i - \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \tag{1.3}$$

[2]Variancia alatt minden esetben szórásnégyzetet fogok érteni.

Mivel feltettük, hogy a hiba nulla várható értékű, így a hiba szórásának becslésében a következő négyzetösszeg szerepel:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip})^2 \tag{1.4}$$

A legkisebb négyzetek elve alapján azt az illesztést fogadjuk el, amelyre a fenti négyzetösszeg minimális. Mivel most feltettük, hogy ε normális eloszlású, ezért a legkisebb négyzetes becslése ebben az esetben éppen meg fog egyezni a maximum likelihood becsléssel, amit a későbbi modellek esetén is előszeretettel fogunk alkalmazni.

1.2.1. A modell gyakorlati alkalmazása

Egy példa segítségével könnyebben megérthetjük a modell működését. Az egyszerűség kedvéért tegyük fel, hogy csak két kategória szerint osztályozzuk a megfigyeléseinket, az egyik a kor, a másik pedig a lakóhely. A könnyebb számolhatóság érdekében kor és lakóhely szerint is csak két-két csoportot különböztetünk meg, legyenek idősek, illetve fiatalok (a gyakorlatban természetesen több csoportot hoznak létre), illetve városiak és vidékiek. Ezek lesznek ennek a változónak (faktornak) a szintjei, tehát a faktor maga a KOR változó, míg a szintjei a *fiatal* és az *idős* lesznek ebben az esetben. Ez a példa így nem igazán élethű, de most csak a modell működését szeretnénk bemutatni, tehát az egyszerűsítéseink elfogadhatók. Ekkor tegyük fel, hogy az átlagos kárnagyságok a biztosító korábbi adatai szerint az alábbiak:

Átl. kárnagyság	*Városi*	*Vidéki*
Fiatal	75.000	62.000
Idős	39.000	21.000

Mivel mind a két faktornak két különböző szintje van, ezért összesen $2 \cdot 2 = 4$ különböző mérési pontunk lesz, mivel a mérési pontok az egyes faktorok egyes szintjeihez tartoznak, ezek pedig X_1 fiatal, X_2 idős, X_3 városi és X_4 vidéki. Ezek $0 - 1$ értéket vesznek fel aszerint, hogy az adott tulajdonsággal rendelkezik-e az adott megfigyelés. Például $X_3 = 1$ azt jelenti, hogy az adott megfigyelés lakóhely szerinti szegmentáció alapján városi. Így a modell feltevéseinknek megfelelően Y-t a következőképpen írhatjuk fel:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \tag{1.5}$$

A négy változónk között lineáris kapcsolat van, ami azt jelenti, hogy nem egyértelmű a felírás. Hiszen gondoljunk csak bele: amennyiben β_1-et és β_2-öt növeljük egy tetszőleges k értékkel, akkor ugyanezt az értéket kivonva β_3-ból és β_4-ből a végső modell ekvivalens lesz az eredetivel. Ez azért áll fenn, mert $X_1 + X_2 = 1$ és $X_3 + X_4 = 1$ is teljesül. Ahhoz, hogy ezt kiküszöböljük, illetve egyszerűbbé tegyük a modellt, az egyik β_i-t elhagyva egyértelműsítjük a felírást. Ekkor a következő egyszerűbb alakot kapjuk:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \tag{1.6}$$

A modellt értelmezhetjük úgy, hogy a magyarázott változó és az egyes magyarázó változók közötti kapcsolatot, a fiatalok esetében β_1, az idősek esetében β_2 írja le, és ezek mellett a modellben megtalálható még egy további additív hatás is β_3, ami a korra való tekintet nélkül csak a városi hatást hordozza. Ha $X_1 = 1$, akkor szükségszerűen $X_2 = 0$ lesz, azaz fiatalt figyeltünk meg, amennyiben $X_1 = 0$, akkor $X_2 = 1$ fog fennállni, azaz idős emberről lesz szó. Ha $X_3 = 1$ akkor városi, ha $X_3 = 0$ akkor pedig vidéki megfigyelésről van szó.

A modellünket még tovább egyszerűsíthetjük, ha kihasználjuk, hogy $X_1 + X_2 = 1$ és azt, hogy mérési pontokról van szó. Ekkor a következő egyszerűbb alakot kapjuk:

$$Y = \beta_1 + (\beta_2 - \beta_1) X_2 + \beta_3 X_3 + \varepsilon \tag{1.7}$$

$$Y = \beta_1 + \beta_2' X_2 + \beta_3 X_3 + \varepsilon \tag{1.8}$$

Ekkor ugyan nem lett kevesebb paraméterünk, tehát valójában nem egyszerűsítettük a modellünket, de ez a felírás nagy jelentősségel fog bírni a bázis szint és annak paraméterbecslése *(intercept term)* kapcsán, amit a későbbiekben tárgyalok a dolgozatomban. Ez egy olyan érték lesz, amely minden β paraméter becslésében meg fog jelenni, és a többi β az ettől való abszolút eltérést hordozza.

Visszatérve a példánkra, az (1.6)-os egyenlettel felírva a feladatunkat, a következő egyenletrendszert kell megoldanunk:

$$\begin{aligned} Y_1 = 75.000 &= \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 1 + \varepsilon_1 \\ Y_2 = 62.000 &= \beta_1 \cdot 1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 + \varepsilon_2 \\ Y_3 = 39.000 &= \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 1 + \varepsilon_3 \\ Y_4 = 21.000 &= \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0 + \varepsilon_4 \end{aligned}$$

A kapott egyenletekből felírhatjuk az (1.4)-es egyenlet alapján a minimalizálandó függvényünket. A β_i-k szerinti parciális deriváltak nullával egyenlő tételével, majd a kapott egyenletrendszer megoldásával minimalizálhatjuk a hibák négyzetösszegét.

$$\sum_{i=1}^{n} \varepsilon_i^2 = (75.000-\beta_1-\beta_3)^2+(62.000-\beta_1)^2+(39.000-\beta_2-\beta_3)^2+(21.000-\beta_2)^2 \quad (1.9)$$

$$\begin{aligned} \tfrac{\partial}{\partial \beta_1} = 0 \quad &\Rightarrow \qquad \beta_1 + \beta_3 + \beta_1 = 75.000 + 62.000 \\ \tfrac{\partial}{\partial \beta_2} = 0 \quad &\Rightarrow \qquad \beta_2 + \beta_3 + \beta_2 = 39.000 + 21.000 \\ \tfrac{\partial}{\partial \beta_3} = 0 \quad &\Rightarrow \quad \beta_1 + \beta_3 + \beta_2 + \beta_3 = 75.000 + 39.000 \end{aligned}$$

A keresett paraméterekre a következő értékeket kapjuk: $\beta_1 = 54.000, \beta_2 = 15.500$ és $\beta_3 = 29.000$. Így a modellünk alapján a becsléseink a következők:

Átl. kárnagyság	*Városi*	*Vidéki*
Fiatal	$\beta_1 + \beta_3 = 83.000$	$\beta_1 = 54.000$
Idős	$\beta_2 + \beta_3 = 44.500$	$\beta_2 = 15.500$

Az illesztett modellünk ezt a becslést adja, ami valamelyest eltér a megfigyelt valós adatoktól, amin nem lepődük meg, hiszen mégis csak modellről van szó. Nem várhatjuk el ennyi korlátozó feltétel (normalitás, stb.) mellett, hogy az teljes mértékben megegyezzen a valósággal. Épp ez a célunk, a sok-sok modell közül kiválasztani azt, ami a legjobb illeszkedést mutatja bizonyos ésszerű feltételek mellett.

1.3. Vektoros jelölés

Ahogy nő a különböző változók és megfigyelések száma, úgy bonyolódik az egyenletrendszer felírása, illetve nehezedik az adatok tárolása, ezért érdemes formulizálnunk a modellünket. A magyarázott Y változót, az előző példa adatait felhasználva, a következő oszlopvektor formájában tároljuk:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 75.000 \\ 62.000 \\ 39.000 \\ 21.000 \end{bmatrix}$$

Legyenek $\underline{X}_1$, $\underline{X}_2$ illetve $\underline{X}_3$ oszlopvektorok, amelyeknek a komponensei jelzik az adott megfigyeléshez kapcsolódó mérési pont értékét. Például $\underline{X}_1$ $i.$ eleme 1, ha az $i.$ megfigyelés fiatalra vonatkozott, és 0 ha idősre. Így pedig:

$$\underline{X}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \underline{X}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \qquad \underline{X}_3 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Legyen $\underline{\beta}$ a paraméterekből álló, míg $\underline{\varepsilon}$ a hibatagokból álló oszlopvektor:

$$\underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

A bevezetett jelölések segítségével a modellünk a következőképpen néz ki:

$$\underline{Y} = \beta_1 \underline{X}_1 + \beta_2 \underline{X}_2 + \beta_3 \underline{X}_3 + \underline{\varepsilon}$$

Ezt pedig még egyszerűbbé tehetjük, ha bevezetjük az $\underline{X}_1$, $\underline{X}_2$ illetve $\underline{X}_3$ oszlopvektorokból álló úgynevezett struktúra *(design)* mátrixot, amit jelöljünk $\mathbf{X}$-szel:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Ezekkel a módosításokkal egyenletrendszerünk a következő általános formát öltötte:

$$\underline{Y} = \mathbf{X} \cdot \underline{\beta} + \underline{\varepsilon}$$

A célunk tehát megkeresni, a $\underline{\beta}$ vektor értékét, ami minimalizálja a hibatagok - $\underline{\varepsilon}$ komponenseiből álló - négyzetösszegét.

1.4. Klasszikus lineáris modell feltevései

Összefoglalva a klasszikus lineáris modell felírását az előző részben bevezetett jelölések segítségével:

$$\underline{Y} = \mathrm{E}[\underline{Y}] + \underline{\varepsilon}, \qquad \mathrm{E}[\underline{Y}] = \mathbf{X} \cdot \underline{\beta}$$

A modellel lapcsolatban a következő feltevésekkel élünk:

(LM1) *Magyarázott változó:* Az $\underline{Y}$ komponensei függetlenek és normális eloszlásúak közös varianciával. A μ_i várható értékek különbözhetnek egymástól.

(LM2) *Lineáris prediktor:* A p darab magyarázó változó lineáris kombinációja a becsülendő β paraméterekkel adja az ún. lineáris prediktort $\underline{\eta}$-t.

$$\underline{\eta} = \mathbf{X} \cdot \underline{\beta}$$

(LM3) *Link függvény:* A kapcsolatot a függő változó és a lineáris prediktor között az ún. link függvény írja le. A lineáris modellben a link függvény az identitás. Ennek a későbbiekben fontos szerepe lesz.

$$\mathrm{E}[\underline{Y}] = \underline{\mu} = \mathbf{X} \cdot \underline{\beta} = \underline{\eta}$$

Nézzük meg milyen feltételezésekkel élünk a modellünkkel kapcsolatban:

1.1. Feltétel (Függetlenség)**.** *Legyen n különböző megfigyelésünk ugyanabban a szegmensben*[3]*, az i-dik megfigyelést jelöljük X_i-vel. Ekkor X_i-k egymástól függetlenek.*

Erre az alapvető feltevésre nem nehéz a gyakorlatban előforduló olyan példát találnunk, ahol ez a feltétel sérül. Vegyünk például egy gépjármű biztosítást, ahol megvan annak a lehetősége, hogy két jármű összeütközik, amelyek ugyanazon biztosító társaságnál biztosítottak, így viszont az ő káruk nem független egymástól. Egy ennél fontosabb példa az úgynevezett katasztrófa károk, amikor nagyszámú biztosítottat károsít meg ugyanaz a természeti katasztrófa, egyazon időben. Ilyen lehet például egy hurrikán, vagy egy jégeső. Az ilyen jellegű károkat, éppen emiatt a biztosítók más típusú modellekkel modellezik.

1.2. Feltétel (Időtől való függetlenség)**.** *Legyen n darab diszjunkt időintervallumunk. Minden X_i essen az i-edik intervallumba. Ekkor az X_i-k egymástól függetlenek.*

Szeretnénk azt feltenni, hogy az időtől független egy kár bekövetkezése, azaz arra gondolunk itt, ha valakinek adott idő alatt, adott számú kára van, akkor például kétszer annyi idő alatt, kétszer annyi kára legyen. Ez persze a gyakorlatban szintén nem teljesül. Például egy lakásbiztosítás esetén, ha valakihez betörnek, akkor valószínűleg ennek hatására riasztót szerel a házába, ami miatt lehet, hogy többet nem mernek oda betörni. Vagy ha valaki a saját hibájából autóbalesetet szenved, a későbbiekben törekedni fog arra, hogy sokkal óvatosabban vezessen. Ennek ellenére ez ésszerű feltételezésnek tűnhet, amely lényegesen leegyszerűsíti a modell építését.

1.3. Feltétel (Homogenitás)**.** *Bármely kettő ugyanahhoz a szegmenshez tartozó, és ugyanakkora kitettséggel (pl. kockázatban töltött idővel) rendelkező megfigyelésünk azonos eloszlást követ.*

[3] Szegmens alatt azonos tulajdonsággal rendelkező megfigyelések halmazát értjük.

A homogenitás sem teljesül minden esetben. Vegyünk például két biztosítottat, akik ugyanakkora kitettséggel rendelkeznek, de különböző évszakban. Télen csúszós utak miatt, nagyobb az esélye egy balesetnek, mint egy nyári hónapban. Ezt kiküszöbölendő a biztosításokat egy éves időtartamra kötik, így gyakorlatilag nem számít, hogy ki, mikor kötötte az adott biztosítást. Így a kitettség szempontájból már homogénnek tekinthetőek a szerződések.

1.5. Lineáris modell korlátozásai

A lineáris modell során egészen könnyen kezelhető problémáink voltak, amelyeket egyszerűen megoldhattunk jól ismert lineáris algebrai módszerekkel. Viszont könnyedén láthatjuk azt is, hogy a megkövetelt feltételezéseket nem könnyű garantálni a modell alkalmazása során.

(i) A megfigyelt adatok normalitására, illetve az közös varianciára tett feltételezés a gyakorlatban általában nem teljesül. (A kárszám például diszkrét eloszlású, ami a nem negatív egészeken van értelmezve. A kárnagyság szintén nem negatív és gyakran jobbra ferde eloszlású.) Amennyiben a magyarázott Y változó nem teljesíti ezeket a feltételezéseket, megkísérelhetjük áttranszformálni a megfigyeléseinket. Vehetjük például a függő változó természetes alapú logaritmusát $\ln(Y)$-t, ami lehet, hogy teljesíti a feltételezéseinket. A probléma abban rejlik, hogy nincs garancia arra, hogy ilyen tulajdonsággal rendelkező függvény a gyakorlatban mindig létezik.

(ii) A magyarázott változó általában pozitív értékeket vesz fel, ami szintén sérti a normalitás feltételezését.

(iii) Az (LM2) illetve az (LM3) feltevésekbe beépített additív hatás, sok esetben nem valóságos, gyakran inkább multiplikatív hatást feltételezünk a magyarázó változók között, ami általában jobban illeszkedik az adatokra.

2. fejezet

Az általánosított lineáris modellek felépítése

Ebben a fejezetben található általánosított lineáris modellel kapcsolatos jelölésrendszer kialakításában az [1]-es szakirodalom nyújtott nagy segítséget. A modell részeinek magyarázatában a [2]-es és a [3]-as könyveket hívtam segítségül, illetve több ízben kellett támaszkodnom magyar és angol nyelvű wikipédián található jegyzetekre is. A fejezetben található 2.1.-es táblázat és 2.2.-es tétel bizonyítása önálló munkám eredménye, a 2.10.-es példa az [1]-es szakirodalomban található tesztadatbázis alapján saját számolásaimat tartalmazza.

2.1. A GLM feltevései

A GLM nem egy konkrét modell, hanem több modell összefoglaló elnevezése, ami az 1.5.-ös fejezetben található korlátozásokat feloldva általánosítja a hagyományos lineáris modellt. Ahelyett, hogy a magyarázott változónak normális eloszlást feltételezünk, annyit teszünk fel, hogy az ún. exponenciális eloszlás-családból származik. Az additív hatás bilincsét pedig a link függvény bevezetésével rázzuk le magunkról. A modellünk az

1.4. fejezet jelöléseit alkalmazva a következőképpen írható fel:

$$\underline{Y} = \underline{\mu} + \underline{\varepsilon}, \quad \text{ahol} \quad \mathrm{E}(\underline{Y}) = \underline{\mu} \quad \text{és} \quad g(\underline{\mu}) = \mathbf{X} \cdot \underline{\beta} \quad \Rightarrow \quad \underline{\mu} = g^{-1}(\mathbf{X} \cdot \underline{\beta})$$

Ez a művelet természetesen pontonként értendő, tehát ezt egy vektor minden egyes elemére végrehajtva egy ugyanolyan dimenziójú vektort kapunk.

(GLM1) *Magyarázott változó:* Az $\underline{Y}$ komponensei függetlenek és az exponenciális eloszlás-családból származnak.

(GLM2) *Lineáris prediktor:* A p darab változó lineáris kombinációja adja az úgynevezett lineáris prediktort $\underline{\eta}$-t.

$$\underline{\eta} = \mathbf{X} \cdot \underline{\beta}$$

(GLM3) *Link függvény:* A kapcsolatot a függő változó és a lineáris prediktor között az ún. link függvény írja le. Ezt jelöljük g-vel, ami legyen monoton és differenciálható, tehát létezzen az inverze is.

$$\mathrm{E}[\underline{Y}] = \underline{\mu} = g^{-1}(\mathbf{X} \cdot \underline{\beta}) = g^{-1}(\underline{\eta})$$

2.2. Az exponenciális eloszlás-család

Teszünk egy kisebb kitérőt, és bemutatjuk ezt az eloszláscsaládot, hiszen ennek ismerete nélkülözhetetlen a GLM teljes megértésének érdekében. A (GLM1) feltevés szerint Y minden megfigyelése az exponenciális eloszlás-családból származik, amit a következőképpen írhatunk fel:

$$Y_i \sim f_i(y_i; \theta_i, \phi) = exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \tag{2.1}$$

$$= exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi, \omega_i) \right\} \tag{2.2}$$

A (2.1)-es képletben $a_i(\phi)$, $b(\theta_i)$ és $c(y_i, \phi)$ az eloszláscsalád tagjától függő függvények, θ_i a várható értékkel kapcsolatos paraméter, ami megfigyelésenként változhat és $\phi > 0$ az úgynevezett skála paraméter, ami a varianciával áll kapcsolatban. Lényeges a kitevőben található $y_i\theta_i$ lineáris kapcsolata. A $b(\theta_i)$ az ún. kumuláns függvény, ami kétszeresen differenciálható, és a második deriváltjának létezik az inverze. Ezt később használni fogjuk. A $c(y_i, \phi)$ függvény nem függ θ_i-től, és nem is lesz jelentősége a továbbiakban számunkra.

Fontos tulajdonsága ezen eloszlásoknak, hogy a várható értéke illetve a szórása, illetve ami ezzel egyenértékű, a θ_i és a ϕ paraméterek egyértelműen meghatározzák az eloszlást a családon belül. Lényeges, hogy Y_i varianciája a várható érték függvénye, amit a következő összefüggéssel írhatunk le:

$$D^2(Y_i) = \frac{\phi \cdot V(\mu_i)}{\omega_i} \tag{2.3}$$

Itt $V(x)$ az úgynevezett varianciafüggvény, ami az eloszláscsalád tagjától függ, ϕ skála paraméter, ω_i pedig az i-edik megfigyeléshez tartozó súly. Ilyen súly lehet például a kockázatban töltött idő vagy a kárdarabszám. Egy fontos tétel a variancia függvény és az exponenciális eloszlás-család tagjai között a következő:

2.1. Tétel. *Az exponenciális eloszlás-családon belül az eloszlást egyértelműen meghatározza a variancia függvény.*

Nézzünk néhány példát az eloszláscsaládból a varianciafüggvénnyel, ha feltesszük hogy $\omega_i = 1$ teljesül.

	$V(x)$	ϕ	$b(\theta_i)$	$c(y_i, \phi, \omega_i)$	$E(Y_i)$	$Var(Y_i)$
Normális	1	σ^2	$\frac{\theta_i^2}{2}$	$-\frac{(y_i\phi)^2}{2\phi^2} - \frac{1}{2}\ln(2\pi\phi^2)$	0	1
Poisson	x	1	$\frac{e^{\theta_i}}{\omega_i}$	$-\ln\left((y_i\omega_i)!\right)$	λ	λ
Gamma	x^2	α	0	$\frac{p\ln(\lambda_i y_i \omega_i)}{\ln(y_i\omega_i\Gamma(p))}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Binomiális	$x(1-x)$	n	$-\ln\left(1 - \frac{e^{\theta_i}}{1-e^{\theta_i}}\right)$	$\ln\binom{\omega_i}{\omega_i y_i}$	p	$p(1-p)$

2.1. táblázat

Fontosnak tartom megjegyezni, hogy a lognormális eloszlás nem az exponenciális eloszlás-családból származik, mégis gyakran alkalmazzák, főként tűzkárok modellezésére, amire vastagfarkú tulajdonsága miatt lehet alkalmas.

2.2.1. A Tweedie modellek

Az ún. Tweedie modellek az exponenciális eloszlás-család skálainvariáns[1] tagjai, amik a következő variancia függvénnyel rendelkeznek:

$$Var(Y) = \phi\mu^p, \tag{2.4}$$

valamely p-re. Nézzük meg p értéke szerint részletezve, hogy pontosan melyik eloszlásról van szó, és minek a modellezésénél vehetjük hasznát:

	Típus	*Eloszlás*	*Mit modellez?*
$p < 0$	folytonos	-	-
$p = 0$	folytonos	normális	-
$0 < p < 1$	nem létezik	-	-
$p = 1$	diszkrét	Poisson	kárgyakoriság
$1 < p < 2$	kevert, nem negatív	összetett Poisson	kárszükséglet
$p = 2$	folytonos, pozitív	Gamma	átlagkár
$2 < p < 3$	folytonos, pozitív	-	átlagkár
$p = 3$	folytonos, pozitív	inverz normális	átlagkár
$p > 3$	folytonos, pozitív	-	átlagkár

Az $1 < p < 2$ esetén az ún. összetett Poisson eloszlásról van szó, ami egy kevert eloszlás, azaz se nem tisztán diszkrét, se nem tisztán folytonos, hiszen pozitív súlyt helyez a nullába, a pozitív számokon pedig folytonos eloszlásként viselkedik. Az eloszlást a következőképpen állíthatjuk elő:

[1] Legyen c pozitív konstans, Y pedig egy adott eloszláscsaládból származó valószínűségi változó. Ekkor azt mondjuk, hogy Y skálainvariáns, ha cY is ugyanabból az eloszláscsaládból származik, mint Y.

Legyen a kárszámunk k, az egyes károkhoz tartozó kárnagyságok pedig rendre $z_1, z_2, \ldots, z_k$, ahol $z_i > 0$ teljesül minden i-re. Ekkor a teljes kárnagyságunk értelemszerűen a következő képlettel számolható:

$$Y = \sum_{i=1}^{k} z_i \tag{2.5}$$

Természetesen $Y = 0$, akkor és csak akkor, ha $k = 0$. Amennyiben k, azaz a kárszám Poisson(λ) eloszlású, a z_i-k, azaz a kárnagyságok pedig független Gamma(α, θ) eloszlásúak, akkor Y összetett Poisson eloszlású valószínűségi változó lesz. Ennek az eloszlásnak, azért van jelentősége, mert közvetlenül alkalmas a kárszükséglet modellezésére. A biztosítónak emellett lehetősége van a kárgyakoriságot és az átlagkárt külön modellezni, majd az így kapott eredményeket összefésülve megkapni a kárszükséglet becslését, én a továbbiakban ezzel az esettel foglalkozom.
A két szélsőséges esetben, ha $p = 1$, akkor Poisson tagszámú, konstans 1 nagyságú kárt adunk össze, míg $p = 2$ esetén egyetlen tagú lesz az összeg, és az egy Gamma eloszlású változó lesz.

2.3. A kárgyakoriság eloszlása

Legyen $N(t)$ egy biztosítási szerződés kárszáma a $[0, t]$ intervallumban, ahol $N(0) = 0$ teljesül. Ekkor az $\{N(t); t \geq 0\}$ sztochasztikus folyamatot kárfolyamatnak hívjuk. Amennyiben a megfigyeléseinket nem szegmentáltuk, és a folyamat teljesíti az 1.2.-es és 1.3.-as feltételezéseinket a folyamatunk Poisson folyamat lesz. Ez motivál minket, hogy feltegyük a kárszámról, hogy Poisson eloszlást követ.
Ennél azonban jobban érdekel minket, hogy a kárgyakoriság, azaz az $Y_i = \frac{X_i}{\omega_i}$ milyen eloszlást követ, ahol ω_i a kockázatban töltött időt jelenti az i-edik megfigyelésre nézve. Nevezzük ezt az Y_i-t ún. relatív Poisson eloszlásnak, ami egy összesűrített Poisson eloszlásnak felel meg, ugyanis az tartója egész számok helyett a racionális számok lesznek.
Kérdés, hogy a gyakorlatban mennyire fordul elő a Poisson eloszlás, azaz mennyire erősek a feltevéseink. A homogenitást például nehéz garantálni, viszont a várható kárgya-

koriság változhat időben, de a kárszám egy év alatt Poisson eloszlást követ (hiszen ekkor $\omega_i = 1$). Ennél komolyabb problémával állunk szemben abban az esetben, amikor a kárszámeloszlásra nem illeszkedik jól a Poisson eloszlás. Tudjuk, hogy minden egyes megfigyelésünk kárszáma Poisson eloszlást követ, viszont minden egyes megfigyelés különböző várahatóértékkel rendelkezik. Ebben az esetben a paramétereknek is valamilyen eloszlást feltételezve, az illesztésünket javíthatjuk. Az így kapott eloszlást keverék Poisson eloszlásnak nevezzük. Részletesebben foglalkozunk ezzel az esettel a 3.1.-es fejezetben.

A kárgyakoriság modellezése során ésszerű feltevés lehet, ha azt szeretnénk, hogy az adataink a következő tulajdonsággal rendelkezzenek: Tegyük fel, hogy minden megfigyelésünk Poisson eloszlást követ. Ebben az esetben, ha két különböző szegmensre ugyanazt a kárszükségletet határoztuk meg, akkor összevonhatjuk ezt a két csoportot. Ekkor az lenne a szerencsés, ha megtartaná az eredeti eloszlást az összevont szegmens is. Szerencsére ez a probléma nem merül fel, ugyanis az összevont csoport kárszáma is Poisson eloszlású lesz, méghozzá ugyanolyan várható értékkel csak nagyobb kitettséggel. Lássuk hogyan: Legyen Y_1 és Y_2 kárszámok, ω_1 illetve ω_2 kitettséggel, és mind a kettő Poisson eloszlást kövessen μ paraméterrel. Ha összevonjuk ezeket a megfigyeléseinket, akkor az új megfigyelésünk a következő lesz:

$$Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2}. \tag{2.6}$$

Mivel $\omega_1 Y_1 + \omega_2 Y_2$ Poisson eloszlást követ $(\omega_1 + \omega_2)\mu$ várható értékkel, tudjuk, hogy Y relatív Poisson eloszlású lesz $\omega_1 + \omega_2$ kitettséggel és ugyanazzal a μ várható értékkel, mint Y_1 és Y_2. Ezen heurisztika alapján mondjuk ki általánosan is az exponenciális eloszlás-családra ezt a tételt:

2.2. Tétel. *Legyen Y_1 és Y_2 független, és tartozzanak ugyanahhoz az exponenciális eloszlás-családhoz, azaz a (2.2)-es jelöléseinket érvényben tartva, rendelkezzenek ugyanazzal a $b(.)$ függvénnyel, ugyanazzal a μ várható értékkel és ugyanazzal a ϕ skálaparaméterrel, de különböző ω_i-vel. Ekkor az ω-val súlyozott átlag $Y = \frac{\omega_1 Y_1 + \omega_2 Y_2}{\omega_1 + \omega_2}$, ugyanahhoz az exponenciális eloszlás-családhoz fog tartozni, ugyanazzal a μ várható értékkel, de $\omega = \omega_1 + \omega_2$*

súllyal.

A bizonyítás megkezdése előtt bevezetjük a momentum-generáló függvényt, majd abból a kumuláns-generáló függvényt, kimondunk két egyszerű állítást, (amelyek bizonyítását a Tisztelt Olvasóra bízzuk), illetve egy lemmát, amiket a 2.1.-es tétel bizonyítása közben fogunk felhasználni.

2.3. Definíció. *A momentum-generáló függvénye, egy exponenciális eloszlás-családból származó eloszlásnak* $M(t) = E(e^{tY})$.

A momentum-generáló függvény n-edik deriváltja a nulla helyen pont az n-edik momentum lesz. Azaz $M^{(n)}(0) = \mathrm{E}(Y^n)$.

2.4. Definíció. *A kumuláns-generáló függvénye, egy exponenciális eloszlás-családból származó eloszlásnak* $\Psi(t) = logE(e^{tY})$.

A kumuláns-generáló függvény első deriváltja a nulla helyen pont a várható érték, míg a második derivált a nulla helyen pont a szórásnégyzet lesz.

$$\Psi'(t) = (logE(e^{tY}))' = \frac{1}{E(e^{Yt})} E(Ye^{Yt}) \Rightarrow \Psi'(0) = E(Y) \tag{2.7}$$

$$(\Psi'(t))' = \left(\frac{E(Ye^{Yt})}{E(e^{Yt})}\right)' = \frac{E(Y^2e^{Yt}) - E(Ye^{Yt})E(Ye^{Yt})}{E^2(e^{Yt})} \Rightarrow \Psi''(0) = D^2(Y) \tag{2.8}$$

2.5. Állítás. *Legyen* c *konstans,* X *pedig az exponenciális eloszlás-család tagja, ekkor* $\Psi_{cX}(t) = \Psi_X(ct)$ *fennáll.*

2.6. Állítás. *Legyen* X *és* Y *függetlenek, és ugyanazon exponenciális eloszlás-család tagjai, ekkor* $\Psi_{X+Y}(t) = \Psi_X(t) + \Psi_Y(t)$ *fennáll.*

2.7. Lemma. *Tegyül fel, hogy* Y_i *exponenciális eloszlás-családhoz tartozik, a (2.2)-es pontban felírt sűrűségfüggvénnyel. Ekkor létezik a kumuláns-generátor függvénye, amit a következő képlettel adhatunk meg:*

$$\Psi(t) = \frac{b(\theta_i + t\phi/\omega_i) - b(\theta_i)}{\phi/\omega_i} \tag{2.9}$$

illetve tudjuk még, hogy:

$$\mu_i = E(Y_i) = \Psi'(0) = b'(\theta_i), \quad D^2(Y_i) = \Psi''(0) = \phi b''(\theta)/\omega_i \tag{2.10}$$

ahol $v(\mu_i)$ a varianciafüggvény, amit a következőképpen fejezhetünk ki: $v(\mu_i) = b''(b'^{-1}(\mu_i))$.

A (2.11)-ben található várható érték és szórásnégyzet a (2.8) illetve a (2.9) egyenletekből jönnek ki, a (2.10)-et felhasználva. A tétel bizonyítása előtt, nézzük meg, hogy a (2.10) egyenlet miként adódik. Először írjuk fel a momentum-generáló függvényét, amit a folytonos eloszlásokra tanult várható érték definíciója szerint kifejtünk. Aztán kiemelünk úgy, hogy az integrál hasában egy másik eloszlás ($\theta' = \theta + t\phi/\omega$ paraméterű hasonló eloszlás-családból származó eloszlás) sűrűségfüggvényét kapjuk, aminek az integrálja 1 lesz. Az eredményünk logaritmusát véve, pedig megkapjuk a kumuláns-generáló függvényt.

$$E(e^{tY}) = \int e^{tY} f_Y(y;\theta,\phi)\mathrm{dy} = \int exp\left\{\frac{y(\theta + t\phi/\omega) - b(\theta)}{\phi/\omega} + c(y,\phi,\omega)\right\}\mathrm{dy} =$$
$$= exp\left\{\frac{b(\theta + t\theta/\omega) - b(\theta)}{\phi/\omega}\right\} \times \int \left\{\frac{y(\theta + t\phi/\omega) - b(\theta + t\phi/\omega)}{\phi/\omega} + c'(y,\phi,\omega)\right\}\mathrm{dy} =$$
$$= exp\left\{\frac{b(\theta + t\theta/\omega) - b(\theta)}{\phi/\omega}\right\}$$

Most térjünk rá a 2.2.-es tétel bizonyítására:

Bizonyítás. Írjuk fel a kapott Y kumuláns-generátor függvényét és kezdjük el átalakítani:

$$\Psi_Y(t) = logE(e^{Yt}) = logE\left(e^{\frac{Y_1\omega_1 + Y_2\omega_2}{\omega_1+\omega_2}t}\right) = logE\left(e^{\frac{Y_1\omega_1}{\omega_1+\omega_2}t} \cdot e^{\frac{Y_2\omega_2}{\omega_1+\omega_2}t}\right) =$$

$$= logE\left(e^{\frac{Y_1\omega_1}{\omega_1+\omega_2}t}\right) + logE\left(e^{\frac{Y_2\omega_2}{\omega_1+\omega_2}t}\right) = \Psi_{Y_1}\left(\frac{\omega_1}{\omega_1+\omega_2} \cdot t\right) + \Psi_{Y_2}\left(\frac{\omega_2}{\omega_1+\omega_2} \cdot t\right)$$

mivel tudjuk, hogy Y_1 és Y_2 exponenciális eloszlás-családba tartozik $b(.)$ függvénnyel, ezt az alakot beírva a 2.7.-es lemmába:

$$\frac{b\left(\theta + t\phi\left(\frac{\omega_1}{\omega_1+\omega_2}\right)/\omega_1\right) - b(\theta)}{\phi/\omega_1} + \frac{b\left(\theta + t\phi\left(\frac{\omega_2}{\omega_1+\omega_2}\right)/\omega_2\right) - b(\theta)}{\phi/\omega_2} =$$

$= \frac{b\left(\theta+\phi\left(\frac{t}{\omega_1+\omega_2}\right)\right)-b(\theta)}{\phi} \cdot \omega_1 + \frac{b\left(\theta+\phi\left(\frac{t}{\omega_1+\omega_2}\right)\right)-b(\theta)}{\phi} \cdot \omega_2 =$

$= \frac{b\left(\theta+\phi\left(\frac{t}{\omega_1+\omega_2}\right)\right)-b(\theta)}{\phi} \cdot (\omega_1+\omega_2) = \frac{b(\theta+t\phi/(\omega_1+\omega_2))-b(\theta)}{\phi/(\omega_1+\omega_2)}$

Tehát ugyanolyan $b(.)$ függvénnyel, várható értékkel és skálaparaméterrel rendelkező eloszlást kaptunk, aminek a súlya a két külön osztály súlyainak összege. □

2.4. A kárnagyság eloszlása

A kárgyakoriság illetve a kárszám becslése mellett a biztosító másik fontos feladata a kárnagyság vizsgálata, hiszen ha ezekre jó becsléssel rendelkezik, akkor az átlagos kárszükségletet is ismeri és ez alapján meg tudja határozni a biztosítás nettó díját. Érdekes kérdés lehet, hogy a kárnagyság esetében, diszkrét avagy folytonos eloszlást érdemes-e illesztenünk. Ha belegondolunk egy kár nagyságát pénzegységben mérjük, - hiszen a biztosítónak abban kell megtérítenie a bekövetkezett kárt - tehát a biztosítási összeg limitéig gyakorlatilag tetszőleges értéket felvehet. Ha viszont azt nézzük, hogy Magyarországon már nem létezik a fillér, így elvben egy kár nagysága csak pozitív egész értékeket vehet fel, a gyakorlatban mégis úgy alakult, hogy abszolút folytonos eloszlásokat illesztenek.
Mivel az előző fejezetben tárgyalt kárgyakoriság esetén multiplikatív Poisson eloszlást feltételeztünk, a kárnagyság esetében multiplikatív Gamma eloszlást fogunk vizsgálni. Ha mind a kettő esetben multiplikatív modellt illesztünk, akkor könnyebb lesz az adatok összefésülése. Legyen az i-dik szegmensbe eső megfigyelés kárdarabszáma ω_i, a kárnagyság pedig Y_i. Nézzük meg ebben az esetben mi a kapcsolat a várható érték és a variancia között. A 2.7.-es lemma alapján $E(Y_i) = \mu_i$ illetve $Var(Y_i) = \phi{\mu_i}^2/\omega_i$-ből kapjuk, hogy:

$$\frac{Var(Y_i)}{E^2(Y_i)} = \frac{\phi}{\omega_i}$$

ami azt jelenti, hogy a relatív szórása konstans az azonos kitettséggel rendelkező ugyanazon szegmensbe tartozó megfigyeléseknek. Másképpen megfogalmazva ez azt jelenti, hogy a szórás arányos a várható értékkel, ami hihetőbb feltételezés, mint ha azt mondanánk hogy az egyes megfigyelések szórása konstans. Gondoljunk bele, ha egy szegmens-

hez tartozó várható érték 10, a szórás pedig 2, akkor egy azonos kitettséggel rendelkező másik szegmens esetén, ahol a várható érték 100, szeretnénk feltételezni hogy a szórás 20 lesz és nem 2.

2.5. A varianciafüggvény és a priori súly

Ahogy a (2.2)-es egyenletben láthattuk, a varianciafüggvényt még további két paraméter határozza meg, a $\phi > 0$ skálaparaméter, és az $\omega_i \geq 0$ priori súly.
A priori súly segítségével adhatjuk meg a modellünkben minden egyes megfigyelés súlyát. A gyakorlatban a modell építésnél, amikor a károk számát vizsgáljuk, nagy jelentősséggel bír minden egyes megfigyelés kockázatnak való kitettsége, azaz nem mindegy hogy egy hónapig, vagy mondjuk egy évig élt egy szerződés. Az eltérő tartam miatt, ezek nem ugyanannyi információt hordoznak, de minden egyes megfigyelés jelentőséggel bír számunkra, ezért a súlyok segítségével -anélkül hogy torzítanánk az eredményeinket- arányosan beépíthetjük ezeket a modellünkbe. A hosszabb kitettséghez, nagyobb súlyt választva, csökken a varianciája az adott megfigyelésnek, így nagyobb befolyással lesz a modellre. Abban az esetben, ha a kárnagyságot vizsgáljuk, akkor a priori súly a kárdarabszám lesz, így a modellünk az átlagos kárnagyságot fogja becsülni.
A skálaparaméter néhány esetben 1, (például Poisson eloszlásnál), így teljesen kiesik a modellből. Általában viszont előre nem ismert az értéke, így becsülnünk kell. A paraméter becslése valójában nem szükséges a GLM megoldásához, de egyes statisztikai értékek (például a sztenderd hiba) meghatározásához szükségünk van rá. A ϕ becslését például maximum likelihood módszerrel végezhetjük el.

2.6. A link függvény

A gyakorlatban történő alkalmazás során felmerülő probléma a klasszikus lineáris modellel, hogy a megfigyelt adatok általában nem teljesítik a normalitás, illetve a konstans szórás feltételét. Ezt orvosolandó két lehetőségünk van. Az egyik ha a megfigyelt adatok egy transzformáltjára próbáljuk alkalmazni a modellt, a másik pedig, ha elhagyjuk

a normalitás feltételét, helyette pedig azt tesszük fel, hogy az eloszlás az exponenciális eloszlás-család egy tagja. Ekkor az Y_i várható értéke helyett, annak valamilyen függvényét közelítjük lineárisan a magyarázó változókkal.

$$E(Y_i) = \mu_i = g^{-1}(\sum_j X_{ij}\beta_i) = g^{-1}(\eta_i) \tag{2.11}$$

Néhány példa a link függvény választására:

	$g(x)$	$g^{-1}(x)$
Identitás	x	x
Logaritmus	$\ln(x)$	e^x
Logit	$\ln(x/(1-x))$	$1/(1+e^{-x})$
Reciprok	$1/x$	$1/x$

Minden eloszláshoz tartozik egy ún. kanonikus link függvény, amellyel a loglikelihood függvény lényegesen egyszerűsödik. A következőket tudjuk:

$$b'(\theta) = \mu \quad illetve \quad g(\mu) = \eta \quad \Rightarrow \quad g(b'(\theta)) = \eta \tag{2.12}$$

Mivel a $b(.)$ és így a $b'(.)$ függvényeket is meghatározza a választott eloszlás, így lehetőségünk van olyan $g(.)$ függvényt választani a modellünkbe, hogy $g(.) = b'^{-1}(.)$ egyenlőség fennálljon, azaz $\theta = \eta$ is igaz lesz. Ez a választás lesz a kanonikus link függvény. Néhány példa a kanonikus link függvényre:

Eloszlás	***Kanonikus link függvény***
normális	identitás
Poisson	logaritmus
Gamma	inverz

A 2.8.-as példában láthatjuk, hogy nem muszáj a kanonikus link függvénnyel számolnunk, de a választása lényegesen megkönnyíti a számolásainkat.

A logit modellt, azaz a logisztikus regressziót akkor alkalmazzák, amikor a magyarázott Y változónk dichotom, azaz két értéket vehet fel. Például a biztosítók ennek segítségével

vizsgálják egy szerződés törlési valószínűségét befolyásoló tényezőket. A törlés ténye indikátor változó, amely 0 értéket vesz fel, ha nem törölték és 1 értéket, amennyiben törölték az adott szerződést. Azt szeretnénk megvizsgálni, hogy milyen megfigyelt tulajdonságok megléte esetén nagyobb a törlési valószínűség. Például a gyakorlatban megfigyelhető, hogy egy havi díjfizetésű szerződő nagyobb valószínűséggel törli a szerződését, mint egy éves díjfizetésű ügyfél.

2.7. Ismert hatás beépítése a modellbe "offset"-ként

Vannak olyan esetek, amikor a magyarázó változó egy ismert hatást tartalmaz, vagy mi azt feltételezzük, hogy ezzel a tulajdonsággal rendelkezik. Ilyenkor célszerű ezt az információt beépítenünk a modellünkbe. Ezt az *offset* segítségével érhetjük el, ami szó szerinti fordításban *eltolás*-t jelent, de mi használjuk inkább a beszédesebb *ismert hatás* elnevezést. Valójában tényleg egy eltolásról van szó, jelöljük az offsetet ξ-vel, ekkor ezt a következőképpen építjük be a lineáris prediktorba:

$$\underline{\eta} = \mathbf{X} \cdot \underline{\beta} + \underline{\xi}$$

Első ránézésre úgy tűnhet, hogy ezzel az eltolással tovább bonyolítottuk a modellünket, de mindjárt meglátjuk, valójában hogyan egyszerűsítettük azt, egyes esetekben.
Multiplikatív GLM-et akarunk illeszteni például a kárdarabszámra, ahol offsetet alkalmazunk. Problémánk lehet az egyes megfigyeléseknek az egymástól eltérő kockázatnak való kitettségével. Gondoljunk csak bele, egy hónap alatt kevesebb a várható kárszám, mint egy év alatt. Ennek a problémának az áthidalásában nagy segítséget nyújt nekünk, ha feltételezünk egy "ismert hatást". Tegyük fel azt, hogy kétszer annyi idő alatt kétszer annyi kára lesz valakinek -azaz a várható kárszám, egyenesen arányos a kockázatnak való kitettséggel-, amibe ha jobban belegondolunk, a valóságtól nem is annyira elrugaszkodott feltételezés.
Nézzük, hogyan egyszerűsödik ezáltal a modellünk. Vegyük a kockázatban töltött idő (ω) természetes alapú logaritmusát, a link függvénynek válasszuk a logaritmust és offsetként

építsük be a modellünkbe. Tehát jelen esetben $\xi = \omega$. Ekkor a következő összefüggéseket írhatjuk fel:

$$E[\underline{Y}] = \underline{\mu} = g^{-1}(\underline{\eta}) = g^{-1}(X.\underline{\beta} + \ln(\underline{\omega})) \tag{2.13}$$

$$E[Y_i] = g^{-1}\left(\sum_j X_{ij}\beta_j + \ln(\omega_i)\right) = exp\left(\sum_j X_{ij}\beta_j + \ln(\omega_i)\right) = exp\left(\sum_j X_{ij}\beta_j\right)\cdot\omega_i \tag{2.14}$$

Láthatjuk, hogy a (2.15)-ös egyenletben az ω additív tagból egy multiplikatív tényező lett.

2.8. A GLM szerkezete

A modellünk, tehát a következő:

$$Y_i = \mu_i + \varepsilon_i \qquad i = 1 \dots n. \tag{2.15}$$

Ahol n a megfigyelések száma, és a következőket tesszük fel:

$$E[Y_i] = \mu_i = g^{-1}\left(\sum_j X_{ij}\beta_j + \xi_i\right) \tag{2.16}$$

$$Var[Y_i] = \frac{\phi V(\mu_i)}{\omega_i} \tag{2.17}$$

ahol:

Y_i	-	a magyarázott változó vektora
$g(x)$	-	a linkfüggvény (egy invertálható, előre ismert függvény)
X_{ij}	-	a faktorokból előállított ún. struktúra *(design)* mátrix
β_j	-	becsülendő paraméterek
ξ_i	-	offset paraméter
ϕ	-	a skálaparaméter
$V(x)$	-	a varianciafüggvény
ω_i	-	a priori súlya az i-edik megfigyelésnek

Az Y_i vektor, a struktúra mátrix, a priori súlyok és a hibatag a megfigyeléseken alapulnak. A link függvény és a varianciafüggvény a modell választásától, azaz a mi döntésünktől függ, a skálaparaméter pedig ismert, vagy becsülhető. Néhány tipikus modell forma:

$\underline{Y}$	**Kárgyakoriság**	**Kárszám**	**Átlagos kárnagyság**
Link függvény	$\ln(x)$	$\ln(x)$	$\ln(x)$
Hibatag	Poisson	Poisson	Gamma
Skála paraméter	1	1	becsült
Varianciafüggvény	x	x	x^2
Priori súly	kitettség	1	kárdarabszám
Offset	0	ln(kitettség)	0

Kárgyakoriság modellezése esetén a kitettséget, mint priori súlyt alkalmazzuk a modellünkben, míg a kárszám modellezése esetén a kitettség logaritmusát ismert hatásként beépíthetjük a modellünkbe, Poisson eloszlás feltételezése mellett. Lásd 2.5.-ös fejezet, ahol részleteztem a kettő közötti lényegi különbséget.

2.9. A bázisszint és a bázis-kárszükséglet

Miután a megfigyeléseinket összegyűjtöttük és szegmentáltuk, minden faktor esetében kitüntetünk egy szintet (általában azt, ahol a legtöbb megfigyelésünk van), és ezt hívjuk bázisszintnek. Ekkor lesz egy úgynevezett bázis-kárszükségletünk *(intercept term)*, ami egy olyan szerződés kárszükséglete, melynek minden tulajdonsága az egyes faktorok bázisszintjének felel meg.

Vizsgáljuk meg az 1.2.-es fejezet utáni példánkat ebből a szempontból. Amennyiben az (1.6)-os egyenletet vizsgáljuk ($Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$), akkor azt mondhatjuk, hogy a modellünkben β_1 írja le a változók közötti kapcsolatot a fiatalok esetében, β_2 az idősek esetében, illetve ezek mellett a modellünkben megtalálható még egy további additív hatás, ami korra való tekintet nélkül csak a városi hatást hordozza.

Amennyiben az (1.7)-es egyenletet vizsgáljuk ($Y = \beta_1 + \beta_2' X_2 + \beta_3 X_3 + \varepsilon$), akkor azt mondhatjuk, hogy a modellünk bázisszintje β_1, ami a városi fiatalok hatását hordozza, emellett két további additív hatás ($\beta_2 - \beta_1$) hordozza az idősek, míg β_3 a vidéki hatást. Tehát, ha idős vidékivel van dolgunk, mind a kettőt hozzáadjuk β_1-hez. Ez azt jelenti, hogy β_1 minden megfigyelésünkre hatással van, emiatt a struktúra mátrix a következő-

képpen fog módosulni:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Tehát ebben az esetben a bázisszintünk a fiatal városiak. Táblázatos formában egy additív modell esetében ez a következőt jelenti:

Faktor szintje	*Paraméter*
Fiatal	0
Idős	β_2'

Faktor szintje	*Paraméter*
Városi	0
Vidéki	β_3

Bázis kárszükséglet:	$+\beta_1$

Mivel itt előre ismert tulajdonságok alapján szegmentáltuk a megfigyeléseinket, ha jön egy új biztosított, könnyedén be tudjuk sorolni az ismereteink alapján és meg tudjuk állapítani a kárszükségletét.

2.10. A modell gyakorlati alkalmazása

Egy példa segítségével könnyebben megérthetjük a modell működését. Az egyszerűség kedvéért térjünk vissza az 1.2.1-es fejezet példájához, és ugyanúgy, mint akkor tegyük fel, hogy csak két kategória szerint osztályozzuk a megfigyeléseinket, az egyik a kor, a másik pedig a terület. Ekkor tegyük fel, hogy az átlagos kárnagyságok az alábbiak szerint alakulnak:

Átl. kárnagyság	*Városi*	*Vidéki*
Fiatal	75.000	62.000
Idős	39.000	21.000

.

A feladat megoldásához meg kell határozzuk az X stuktúra mátrixot, illetve a $\underline{\beta}$ paramétervektort. Választanunk kell egy link függvényt, valamilyen hibataggal.

A példánkban Gamma eloszlást feltételezünk log link függvénnyel. Láthatjuk, hogy nem a Gamma eloszláshoz tartozó kanonikus link függvényt (inverz) választottam, ami lényegesen megkönnyítené a számolást, de ennek ellenére is eredményre fogunk jutni. Amennyiben 3 kategóriára bontanám fel a kor és a lakóhely szerint is a megfigyeléseinket, akkor már 5 paraméterünk lenne, amit a kapott egyenletrendszerből a kanonikus link függvény nélkül nem tudnánk mechanikusan megoldani.

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 75.000 \\ 62.000 \\ 39.000 \\ 21.000 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Az általam megoldott példában a link függvény alapján a következőt írhatjuk fel:

$$E(\underline{Y}) = g^{-1}(X \cdot \underline{\beta}) = \begin{bmatrix} g^{-1}(\beta_1 + \beta_3) \\ g^{-1}(\beta_1) \\ g^{-1}(\beta_2 + \beta_3) \\ g^{-1}(\beta_2) \end{bmatrix} = \begin{bmatrix} e^{\beta_1 + \beta_3} \\ e^{\beta_1} \\ e^{\beta_2 + \beta_3} \\ e^{\beta_2} \end{bmatrix}$$

A Gamma eloszlás sűrűségfüggvénye:

$$f(x; \mu, \phi) = \frac{x^{-1}}{\Gamma\left(\frac{1}{\phi}\right)} \cdot \left(\frac{x}{\mu\phi}\right)^{\frac{1}{\phi}} \cdot e^{-\frac{x}{\mu\phi}} \tag{2.18}$$

Ebből a log-likelihood függvényt kiszámolva a következőt kapjuk:

$$L(x; \mu, \phi) = \sum_{i=1}^{n} \frac{1}{\phi}(\ln \frac{x_i}{\mu_i} - \frac{x_i}{\mu_i}) - \ln x_i - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \tag{2.19}$$

A log link függvény $\mu_i = exp(\sum_j X_{ij}\beta_j)$ behelyettesítésével a kötvekezőt kapjuk:

$$l(x, e^{X\beta}, \phi) = \sum_{i=1}^{n} \frac{1}{\phi}(\ln x_i - \sum_{j=1}^{p} X_{ij}\beta_j - \frac{x_i}{exp(\sum_{j=1}^{p} X_{ij}\beta_j)}) - \ln x_i - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \tag{2.20}$$

Kibontva a szummákat, a feladat szerint a következő egyenletet kapjuk:

$$\begin{aligned} l(x, \mu, \phi) \quad = \quad & \frac{1}{\phi}(\ln 75.000 - (\beta_1 + \beta_3) - \frac{75.000}{e^{(\beta_1+\beta_3)}}) - \ln 75.000 - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \\ & + \frac{1}{\phi}(\ln 62.000 - \beta_1 - \frac{62.000}{e^{\beta_1}}) - \ln 62.000 - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \\ & + \frac{1}{\phi}(\ln 39.000 - (\beta_2 + \beta_3) - \frac{39.000}{e^{(\beta_2+\beta_3)}}) - \ln 39.000 - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \\ & + \frac{1}{\phi}(\ln 21.000 - \beta_2 - \frac{21.000}{e^{\beta_2}}) - \ln 21.000 - \frac{\ln \phi}{\phi} - \ln \Gamma\left(\frac{1}{\phi}\right) \end{aligned}$$

Kihasználva, hogy a β-tól független tagok nem befolyásolják a maximum értékét, az egyenletet leegyszerűsítve a következő alakra jutunk:

$$l^*(x, e^{X\beta}) = -2(\beta_1+\beta_2+\beta_3) - 75.000e^{-(\beta_1+\beta_3)} - 62.000e^{-\beta_1} - 39.000e^{-(\beta_2+\beta_3)} - 21.000e^{-\beta_2} \tag{2.21}$$

A kapott egyenlet maximalizálásához határozzuk meg a β_i-k szerinti parciális deriváltakat.

$$\frac{\partial l^*}{\partial \beta_1} = 0 \quad \Rightarrow \quad 2 = 75.000e^{-(\beta_1+\beta_3)} + 62.000e^{-\beta_1} \tag{2.22}$$

$$\frac{\partial l^*}{\partial \beta_2} = 0 \quad \Rightarrow \quad 2 = 39.000e^{-(\beta_2+\beta_3)} + 21.000e^{-\beta_2} \tag{2.23}$$

$$\frac{\partial l^*}{\partial \beta_3} = 0 \quad \Rightarrow \quad 2 = 75.000e^{-(\beta_1+\beta_3)} + 39.000e^{-(\beta_2+\beta_3)} \tag{2.24}$$

Ezután egyenlővé téve a (2.22) és (2.24) illeve a (2.23) és (2.24) egyenleteket, majd ezek természetes alapú logaritmusát véve és egymásból kivonva megkapjuk β_3-at, majd azt visszahelyettesítve (2.22)-be megkapjuk β_1-et illetve (2.23)-ba helyettesítve β_2-t. A keresett paraméterekre a következő értékeket kapjuk:

β_1	6,3283
β_2	5,4600
β_3	0,4047

.

Ezekkel továbbszámolva a következőképpen fog alakulni a táblázatunk:

Átl. kárnagyság	*Városi*	*Vidéki*
Fiatal	84.000	56.000
Idős	35.200	23.500

.

Az illesztett modellünk ezt a becslést adja, ami ismét eltéréseket mutat a megfigyelt valós adatoktól. Dolgozatom terjedelmi korlátai miatt nem számoltam ki különböző eloszlások feltételezése esetén, hogy melyik mennyire közelíti az adatainkat, most csak a számolás mikéntjét akartam bemutatni, nem a legjobb modellt megkeresni. Természetesen másik eloszlás feltételezésével, illetve eltérő link függvény használatával más és más eredményeket kaphatunk.

3. fejezet

A modell alkalmazhatósága

Ebben a fejezetben a [2]-es és a [3]-as és a [4]-es szakirodalmak voltak nagy segítségemre. A 3.3.-as részben található példák saját ötleteimet tartalmazzák.

3.1. A „túlszórás”

A Poisson eloszlás egyik lényeges tulajdonsága, hogy a várható értéke és a szórásnégyzete megegyezik. Legyen X Poisson eloszlású valószínűségi változó λ paraméterrel, ekkor teljesül, hogy:

$$E(X) = D^2(X) = \lambda$$

A 2.3.-as fejezetben arról volt szó, hogy a Poisson eloszlás jó illeszkedést mutat a kárgyakoriság modellezésekor, bár vannak olyan esetek, amikor keverék Poisson eloszlást érdemesebb alkalmaznunk.

3.1. Definíció. *Az X valószínűségi változó keverék Poisson eloszlást követ, ha létezik olyan Y valószínűségi változó, hogy X-nek az Y-ra vonatkozó feltételes eloszlása, Y paraméterű Poisson eloszlás.*

A gyakorlatban számos esetben nem teljesül, hogy a megfigyeléseink várható értéke és szórása megegyezik. A variancia gyakran meghaladja az átlagot. A szakirodalomban ezt a jelenséget hívják *„overdispersion”*-nek, azaz *„túlszórás”*-nak. A túlszórás hátterében általában az alábbi két ok állhat:

1. Az egyik ilyen oka a gyakorlatban gyakran előforduló túlszórásnak az, hogy a megfigyeléseink alkalmával lesznek olyan magyarázó változók, amelyek kimaradnak az elemzésből. Ez azért történhet meg, mert egész egyszerűen megfeledkezünk róluk, vagy azért, mert nem tudunk róluk adatot gyűjteni. Például gépjármű felelősségbiztosítás esetében vizsgáljuk a kárszámot, amit a vezető kora és a lakhelye befolyásol, mi azonban e két tényező közül csupán a vezető kora szerint végezzük az elemzésünket. Ebben az esetben, ha rögzítjük a vezető korát, akkor a modellünk szerint ehhez a szinthez tartozó biztosítottak homogének lesznek, azaz függetlenek lesznek a lakóhelyüktől. Viszont minden egyes lakóhely szerint eltérő átlagaink lehetnek, azaz a várható érték nem lesz konstans, ahogyan azt a Poisson eloszlás feltételezné. Ennek következtében a károk szórásnégyzete nagyobb lesz annál, mint amit a Poisson eloszlás esetében várnánk.

2. A másik ilyen ok lehet a túlszórásra, hogy a megfigyelt adataink nem függetlenek egymástól. Ha a gyakorlatban ez nem teljesül (lásd: 1.4.-es következmény), akkor a Poisson eloszlás alapján a vártnál nagyobb lesz a szélső értékek előfordulási gyakorisága, ennek eredményeként pedig emelkedik a változó szórása, ami sérti a Poisson eloszlás azon alapfeltevését, hogy a várható értéke és a szórásnégyzete megegyezik.

A probléma a gyakorlatban előforduló túlszórással az, hogy a becsült együtthatók standard hibái a ténylegesnél kisebbek lesznek, aminek egyenes következménye, hogy a szignifikancia szintek a valóságosnál kedvezőbb képet mutatnak. Ennek az oka, hogy a várható értéket és a szórásnégyzetet azonosnak vesszük, holott az utóbbi valójában nagyobb. A probléma feloldására két lehetőségünk van:

1. Az egyik a standard hibák utólagos kiigazítása, korrigálása. Ennek a lényege, hogy a standard hibákat a túlszórás mértékét jelző paraméter négyzetgyökével szorozzuk.

2. A másik számunkra érdekesebb megoldás, keverék Poisson eloszlás alkalmazása, azaz a várható értékét nem konstansnak tekintjük, hanem valamilyen eloszlást feltételezünk neki. A leggyakoribb, és matematikai szempontból a legcélszerűbb feltételezés az, ha a várható értéknek Gamma eloszlást feltételezünk. Ez azért kellemes számunkra, mert ez a keverék Poisson eloszlás az ún. negatív binomiális eloszlás lesz, ami amellett, hogy megkönnyíti a számolásainkat az illesztésünket is jelentősen javíthatja. Ezzel az a probléma, hogy a Gamma eloszlás feltételezését nem magyarázza - az egyszerűsödő számoláson kívül - semmi, nyugodtan használhatnánk például lognormális eloszlást is, de azzal elég kellemetlen lenne a további számolás.

3.2. Modell választás

Az előző fejezetben többek között arról is szó volt, hogy kevesebb magyarázó változó használata túlszóráshoz vezethet. A kérdés az, hogy mennyi az optimális paraméterek száma. Kijelenthetjük-e azt, hogy minél több paramétert használunk, annál jobban fog illeszkedni a modellünk? Amennyiben maximális paraméter számot használunk, az illesztésünk jobb lesz, de emlékezhetünk a „A gépjármű és a vezető kora” című részre a dolgozat elején, ahol négy különböző faktorunk volt, de a maximális négy paraméter használata helyett, jobbnak láttuk ha csak három paramétert vezetünk be.
Valójában nem olyan modellt keresünk, ami tökéletesen illeszkedik az adatainkra, sokkal inkább olyat, ami a lehető legjobban előrejelzi a következő év kimenetelét a biztosító számára fontos adatok szempontjából. Ez azt jelenti, hogy olyan változókat akarunk bevonni a vizsgálatba, amik statisztikailag szignifikáns voltuk mellett az időben stabilnak tekinthetőek a nettó kárszükséglet meghatározásakor. Minél nagyobb a modellünk, annál nagyobb a kockázata annak, hogy a különböző változók között korreláció lesz és a modellünknek nem lesz tényleges prediktív hatása. Fontos megjegyeznünk, hogy lehet

egy változó statisztikailag szignifikáns a kárszükséglet meghatározásának szempontjából, anélkül hogy szignifikáns vagy fontos lenne, a szó hétköznapi jelentésében.

Minden egyes, a modellünkhöz hozzáadott magyarázó változóval javíthatjuk az illesztésünket. Ugyanakkor minden egyes indokolatlanul hozzáadott változó feleslegesen bonyolítja a paraméterbecslésünket. Érezhető, hogy ebben az esetben kompromisszumos megoldásra van szükség, a modellünkbe beválogatott változók számával kapcsolatban.

Bemutatunk két olyan kritériumot, ami segít megtalálni az egyensúlyt a beválasztott paraméterek száma és a modell alkalmazhatósága között. Az egyik ilyen „Akaike információs kritériuma" *(Akaike's Information Criterium)* a másik pedig a „bayesi információs kritérium" *(Bayesian Information Criterium)*.

$$AIC \equiv -2l + 2p, \quad BIC \equiv -2l + p \ln n.$$

Ezekben a kifejezésekben p a paraméterek számát, n pedig a megfigyelések számát jelöli, míg l a log-likelihood függvény számunkra lényeges része: $l = \frac{S}{\sigma^2}$, ahol $S = \sum_i (y_i - \hat{y_i})^2$, azaz az eltérések négyzetösszege.

A log-likelihood függvényünket maximalizálni akarjuk, minél nagyobb az l értéke (úgy lesz kisebb $-l$), annál pontosabb az illesztésünk, azaz annál kisebb lesz az S értéke. Újabb paraméterek hozzávételével növelhetjük l értékét, ezért a mind a két kritérium egy a paraméterek p számától függő büntető kifejezéssel látja el a modellt. Láthatjuk, hogy ez a büntető tag a BIC-nél nagyobb, mint az AIC-nál.

Láthatjuk, hogy mind a két kritérium, az eltérések négyzetösszegét veszi, és ahhoz adja hozzá még valamilyen formában a paraméterek számát. Azt, hogy konkrét esetben melyiket használjuk, a megfigyelések száma szokta eldönteni, amennyiben n nagy, érdemesebb a bayesi kritériumot használni, míg kisebb megfigyelés szám esetén inkább Akaike kritériumát. Amikor két modell jóságát akarjuk összvetni, először eldöntjük, hogy melyik mutatószám alapán szeretnénk az összehasonlítást elvégezni (AIC vagy BIC), ezután összehasonlítjuk a kapott értékeket és a kisebbel rendelkező modellt választjuk.

3.3. A változók közötti kapcsolat

Az első fejezetben szó volt arról, hogy az alkalmazott modellek egyik elvárt tulajdonsága, hogy a változók közötti kapcsolatot kezelje. Fontos ismernünk ezeket, hiszen ismeretük birtokában pontosíthatjuk, és ezáltal megbízhatóbbá tehetjük a modellünket. A biztosítási matematikában, amikor pédául a kárszükségletet szeretnénk meghatározni, próbáljuk a lényeges változókat bevonni a vizsgálatunkba, viszont minél több változóval dolgozunk, annál nagyobb az esélye, hogy lesznek olyanok, amik között valamilyen kapcsolat áll fenn. Két lényeges kapcsolatot mutatok be, melyek megértése és a kettő megkülönböztetése nagyon fontos feladata a biztosító aktuáriusának.

1. Az egyik ilyen gyakran előforduló kapcsolat a változók között a korreláció. Nézzünk egy példát erre. Vizsgáljuk a kockázatban töltött időt, egy lakásbiztosítás esetén, ahol két szempontot veszünk figyelembe; a ház típusát és elhelyezkedését:

Kockázatban töltött idő	*Családi ház*	*Társasház*
Vidék	20 év	765 év
Nagyváros	234 év	34 év

   Jól látható, hogy társasházak esetén vidéken, míg családi házak esetén nagyvárosban lesz a kockázatban töltött idő jelentősebb, azaz a társasház a vidékkel, míg a családi ház a nagyvárossal korrelál.

2. A másik ilyen kapcsolat a változók között az interakció, melyről akkor beszélünk, amikor az egyik faktor valamely szintjén hordozott hatás a modellünkben attól függ, hogy a másik változó milyen szintjén tartózkodunk. Nézzünk erre is egy példát. Egy gépjármű felelősségbiztosítás esetén megfigyelték, hogy a fiatal férfiaknak jóval nagyobb a kárnagyságuk, mint a fiatal nőknek, míg amikor középkorú férfiakat vizsgálták kárnagyság szempontjából, már kisebb volt az a nőkkel szemben, sőt ha az időseket is vizsgálták, már a nőknek volt nagyobb a kárgyakoriságuk. Ezt hívják tehát interakciónak, azaz nem mindegy, hogy milyen szinten (milyen életkorban) vizsgáljuk a férfiak és a nők egymáshoz viszonyított kárnagyságát.

Átlagkár (eFt)	*Férfi*	*Nő*
Fiatal	91	30
Középkorú	49	63
Idős	21	100

Vizsgáljuk meg az általam készített interakciót (az R program segítségével), hogy tudjuk beépíteni a modellünkbe. Mivel a nemek hatása függ attól, hogy kor szerint milyen szinten vagyunk, ezért ezt kezelnünk kell a modellünkben, amit egy új változó bevezetésével tudunk megoldani, ami összesen $2 \cdot 3 = 6$ különböző szinttel rendelkezik, amelyek a következők lesznek: *Fiatal:Férfi*, *Fiatal:Nő*, *Középkorú:Férfi*, *Középkorú:Nő*, *Idős:Férfi* és *Idős:Nő*, ezeket pedig szintenként már kezelni tudjuk.

Az R program segítségével GLM-et illesztem az adatokra, először Gamma eloszlás majd normális eloszlás feltételezésével, mind a két esetben log link függvénnyel és ugyanazt az eredményt kaptam a paraméterek becslésére, csak a sztenderd hibákban voltak eltérések:

Változó	**Paraméter**	**Exp(Paraméter)**
intercept	4,51	91
Férfi	0,00	1,00
Fiatal	0,00	1,00
Idős	-1,45	0,23
Középkorú	-0,60	0,55
Nő	-1,10	0,33
Idős:Nő	2,65	14,15
Közpkorú:Nő	1,34	3,81

Ez pedig azt jelenti, hogy egy új biztosított esetén, a kora és a neme alapján az alábbi szorzók közül kell alkalmazni azt, amelyik szegmensbe ő tartozik. Láthatjuk, hogy a kor szerint a fiatalok, a nem szerint pedig a férfiak hiányoznak a táblázatból, mégpedig azért mert ez a szint lesz lesz a bázisszint vagy intercept. A fiatal férfiak átlagkára tehát 91 eFt lesz, ha viszont egy idős férfit vizsgálunk, akkor már $91 \cdot 0,23 = 21$ eFt lesz az átlagkár, míg például egy középkorú nő esetén $91 \cdot 0,55 \cdot 0,33 \cdot 3,81 = 63$ eFt lesz az átlagos

kárnagysága, ahol a 91 az intecept, a 0,55 a középkorúak hatása, a 0,33 a nők hatása illetve 3,81 a középkorú nők hatása. Ezek a szorzók (a 14,15 és a 3,81) tehát egy plusz hatást hordoznak, amik a változók (a nem és a kor közötti) közötti interakciót korrigálják.

3.4. A nagy károk

A biztosító lényeges feladata, az esetlegesen felmerülő nagy károk modellezése illetve kezelése. Ezek, habár ritkán fordulnak elő, elég nagyok ahhoz, hogy a becslésünket torzíthassák, ezért érdemes a hatásukat valahogy csökkenteni. Az egyik módszer az ilyen jellegű nagy károk kezelésére, ha úgymond megcsonkítjuk az adatainkat, azaz kinevezünk egy küszöböt, aminél nagyobb károkat nem a tényleges mértékén veszünk figyelembe. Jelöljük a kárnagyságot X_k-val, ekkor bevezetünk egy új változót mégpedig: $\tilde{X}_k = min(X_k, c)$, ahol $c \in \mathbb{R}^+$ a bevezetett küszöb. Természetesen a küszöb bevezetésével az egyes károk értékét módosíthatjuk, hogy könnyebb legyen a kárnagyság modellezése, de az átlagos kárnagyságot illetve a tényleges összkárt nem változtathatjuk meg, hiszen az az adat fontos nekünk a nettó kárszükséglet meghatározásakor. Tehát ez a módszer két kérdést vet fel számunkra, hogyan válasszuk meg c-t, illetve mi legyen a nagy károknak a küszöb fölé eső részével, amelyek befolyásolják az átlagos kárnagyságot.

A c küszöb megválasztásánál két dolgot kell figyelembe vennünk. Próbáljunk minél nagyobb küszöböt választani, annak érdekében, hogy az elemzésünk releváns legyen, ugyanakkor próbáljunk minél kisebb küszöböt választani, azt a célt szem előtt tartva, hogy az elemzésünk ne torzítson.

A küszöb fölé eső kárrésszel is kell valamit kezdenünk, azt is bele kell építenünk a díjba, hiszen egy ilyen kár bekövetkezése esetén is fizetnie kell a biztosítónak. Erre egy megoldás, ha ezt a kárszükségletet például az elszenvedett károk arányában szétosztjuk a kárt okozó biztosítottak között, így a nagy károkat mindenki megfizeti, de nem olyan mértékben, mintha egyénileg kellene ezzel a lehetőséggel számolnia.

4. fejezet

A credibility elmélet

A credibility elmélet megértésében és elsajátításában a [3]-as és az [6]-os szakirodalmak nyújtottak nagy segítséget. Ebben a fejezetben található példa adatai fiktívek míg a megoldása saját számolásaim eredménye. Az algoritmusok felvázolásánál az [4]-es szakirodalomra támaszkodtam.

4.1. A multifaktor

Manapság egyre nagyobb a verseny a biztosítási piacon, így egy újonnan megalakult biztosító számára nem egyszerű feladat a kezdeti ügyfélportfólió kialakítása, aminek egyenes következménye, hogy nehezen tud saját adatokat gyűjteni, így a különböző statisztikák meghatározásához kénytelen felhasználni a publikusan (Magyarországon például a MABISZ vagy a KSH oldalán) fellelhető adatokat. Felmerül ilyenkor a kérdés, milyen arányban támaszkodhat a saját adataira, és milyen arányban a publikusan fellelhető, esetleg más biztosítók adataira.

Ha nem rendelkezünk elegendő adattal, azaz vannak olyan szegmensek, ahol kevés megfigyelés áll rendelkezésünkre, ott összevonhatunk különböző szinteket, ezzel hihetőbbé téve a becslésünket, viszont erre nem mindig van lehetőségünk. Az olyan faktorok ese-

tén, amiknek túl sok szintje van (ún. *multilevel factor*-ok), előfordulhat hogy szinte egyik szinten sincs elegendő megfigyelésünk. Ilyen például egy gépjármű felelősségbiztosítás alkalmazása esetén, az irányítószám szerinti csoportosítás, amikor több ezer különböző szintünk lehet. Ebben az esetben is alkalmazhatnánk, a különböző szintek összevonását, mondjuk az egymással szomszédos területek alapján, de ekkor valószínűleg nem cselekednénk helyesen, hiszen egy nagyváros és a szomszédságában fekvő kis tanya, kockázati besorolás szempontjából aligha vannak egy szinten. Valószínűbb, hogy az összevonáshoz még további adatokra lesz szükségünk az adott településekről, például átlagos jövedelem, népsűrűség vagy átlagos életkor, amiknek a megállapításához további kutatómunkára lenne szükség. Az irányítószám szerinti csoportosításhoz hasonlóan, amikor az autók típusa alapján akarunk szegmenseket létrehozni, hasonló akadályba ütközünk.
A közös ebben a fent említett két esetben, hogy mind a két változó kategorikus, és hogy több szinten is nagy valószínűséggel kevés megfigyelésünk van ahhoz, hogy megbízható becslést tudjunk adni az adott kategóriákra. Akkor is MLF[1]-ről beszélünk, amikor vannak olyan szegmensek, ahol megfelelő számú megfigyelés áll a rendelkezésünkre, de ez nem minden szinten valósul meg. Ilyen esetekben hívhatjuk segítségül a credibility-elméletet és javíthatjuk vele a becslésünket.

4.2. A credibility elmélet

A credibility elmélet gyakorlatilag nem más, mint a biztosítási matematikai bayesi megközelítése. Legyen például Y a kárkifizetés, az egyes megfigyeléseinket jelöljük Y_{jt}-vel, ω_{jt} kitettséggel, ahol j a multifaktor szintje és t jelentse a szinten belüli megfigyelés sorszámát, ami persze függhet j-től, tehát valójában $t = t(j)$, de a jelölés megkönnyítése miatt ettől most eltekintünk. Ekkor legyen $\overline{Y}_{j.} = (\sum_t \omega_{jt} Y_{jt})/(\sum_t \omega_{jt})$, azaz a súlyozott átlag. Az alapötlete a credibility elméletnek, hogy vegyük az $\overline{Y}_{j.}$-t az adott j szinten és az egész portfólió várható értékét μ-t és írjuk fel az alábbi lineáris kombinációjukat:

$$z_j \overline{Y}_{j.} + (1 - z_j)\mu, \tag{4.1}$$

[1]MLF - Multilevelfactor

ahol μ jelentse a teljes súlyozott átlagot, azaz:

$$\overline{Y}_{..} = \frac{\sum_j \omega_{j.} \overline{Y}_{j.}}{\sum_j \omega_{j.}}, \qquad \omega_{j.} = \sum_t \omega_{jt} \tag{4.2}$$

ami stabilabbnak tekinthető $\overline{Y}_{j.}$-nél.

A $0 \leq z_j \leq 1$-t hívjuk credibility faktornak (vagy súlynak), ami azt mutatja meg, hogy milyen arányban vesszük figyelembe az adott szint saját adatait. Ha $z_j = 1$ akkor teljes credibility becslésről beszélünk. Kérdés, hogyan érdemes megválasztanunk z_j-t. Ha mondjuk a multilfaktor minden szintje különböző adathalmazhoz tartozik, akkor úgy lenne érdemes megválasztani z_j-t, hogy a több megfigyeléssel rendelkező esetben legyen nagyobb hiszen ekkor megbízhatóbb becslést várunk az adatokból, míg kevesebb megfigyelés esetén kevésbé jelentős, tehát a kitettséggel arányosnak kell lennie.

Egy példa

Vizsgáljunk egy GFB adatállományt bonus-malus rendszerrel több évre visszamenőleg. Jelöljük ω_{jt}-vel egy-egy szerződés kitettségét, ahol j jelöli, hogy melyik szerződőről van szó, t pedig azt, hogy melyik naptári évről van szó. Legyen N_{jt} a megfelelő időszakban okozott károk száma, ekkor $\omega_{j.} = \sum_t \omega_{jt}$ jelöli a teljes kitettséget, míg $N_{j.} = \sum_t N_{jt}$ az összes megfigyelt kárszámot az adott biztosítottra. Ezekkel a jelölésekkel a tapasztalati kárgyakoriság $\overline{Y}_{j.} = N_{j.}/\omega_{j.}$ lesz.

Ésszerű feltevés, hogy N_{jt} kövessen Poisson eloszlást, $\omega_{jt}\lambda_j$ paraméterrel (λ_j a $j.$ szerződőre vonatkozó paraméter, az adott biztosított egy év alatt várható kárainak száma, ismeretlen számunkra, ezért feltételezünk rá egy eloszlást), ekkor $N_{j.} \sim Poi(\omega_{j.}\lambda_j)$. Azért szerencsés a paramétert így választanunk, mert ekkor $E(\overline{Y}_{j.}) = \lambda_j$ teljesül, azaz várható kárgyakoriság a $j.$ szerződésre pont λ_j lesz.

Korábban már volt róla szó, hogy érdemes magát a paramétert is valószínűségi változónak tekinteni, azaz tegyük fel λ_j-ről, hogy egy tetszőleges Λ_j eloszlást követ. A különböző Λ_j-kről annyit tételezünk fel, hogy egymástól függetlenek és azonos eloszlásúak. Az, hogy λ_j-nek is egy eloszlást feltételezünk azt jelenti, hogy az egyéneket megkülönböz-

tetjük, azaz nem feltételezzük azt, hogy mindenki ugyanolyan várható értékkel okoz balesetet, mindenkinek egyénre szabott kárgyakorisága lesz. Ha belegondolunk, ez tényleg egy ésszerű feltevés. A biztosító egy új termék bevezetésénél nem ismeri λ_j-t, a kártapasztalata pedig még nem elég nagy ahhoz, hogy hihető becslést adjon neki. Ekkor jön jól számukra a bónusz rendszer, hiszen ekkor minden egyes biztosítottnak, egyénre szabottan befolyásolja az elmúlt időszakban okozott károinak a száma a következő időszaki bónusz besorolását, azaz a biztosítás díját. Ha valaki kárt okoz, alacsonyabb szintre sorolják és magasabb díjat fizet, ha valaki kármentes magasabb szintre kerül és alacsonyabb díjat fizet. A bónusz rendszer nagy előnye még, hogy a biztosítónak lesz előzetes információja egy lehetséges új szerződőről, hiszen a bónusz fokozatát mindenki viszi magával, egy esetleges biztosítóváltás alkalmával.

A bónusz rendszer bevezetésének segítségével nem kell ismernünk λ_j-t, viszont Λ_j-t sem ismerjük, de az összes megfigyelésünk alapján azt tudjuk, hogy $E(\Lambda_j) = \mu$. Egy biztosított kárszáma az adott időszakban a korábbi jelölésünk alapján $N_{j.}$, ekkor a következőre vagyunk kíváncsiak: $E(\Lambda_j|N_{j.})$, azaz amennyiben a biztosított várható kárgyakorisága Λ_j eloszlást követ, és feltételezzük, hogy az időszakban $N_{j.}$ kárszáma van, akkor mennyi lesz a várható kárgyakorisága. Ha feltesszük, hogy Λ_j Gamma eloszlást követ α és β paraméterekkel, akkor $\mu = \alpha/\beta$ és fennáll a következő összefüggés:

$$E(\Lambda_j|N_{j.}) = z_j \frac{N_{j.}}{\omega_{j.}} + (1 - z_j)\mu, \quad ahol \qquad z_j = \frac{\omega_{j.}}{\omega_{j.} + \beta} \tag{4.3}$$

A felírt összefüggés egy jól ismert állítás következménye, ami az a priori és az a posteriori Gamma eloszlás várható értékei közötti kapcsolatot írja le. Azt tudjuk, hogy az a priori eloszlás várható értéke α/β, ami az a posteriori esetben N_j-től és ω_j-től is függeni fog a következőképpen:

$$\frac{\alpha}{\beta} \to \frac{\alpha + N_j}{\beta + \omega_j} \tag{4.4}$$

Ebből pedig már könnyen megkapjuk a (4.3)-as összefüggést:

$$\frac{\alpha + N_j}{\beta + \omega_j} = \frac{\alpha}{\beta + \omega_j} + \frac{N_j}{\beta + \omega_j} = \frac{\alpha}{\beta}\frac{\beta}{\beta + \omega_j} + \frac{N_j}{\omega_j}\frac{\omega_j}{\beta + \omega_j} = (1 - z_j)\mu + z_j \frac{N_j}{\omega_j} \tag{4.5}$$

Ha $\overline{Y}_{j\cdot} = N_{j\cdot}/\omega_{j\cdot}$, akkor az itt kapott állítás megegyezik a (4.1)-es egyenlettel, a különbség annyi csupán, hogy meg tudtuk becsülni z_j-t. Mivel Λ_j-nek Gamma eloszlást feltételeztünk, a megfigyeléseinkből meg tudjuk becsülni az α és a β paramétereket is, amiből pedig meg tudjuk határozni a credibility faktort is. Minél nagyobb az $\omega_{j\cdot}$, annál nagyobb lesz a z_j, azaz annál nagyobb súllyal fog latba esni az egyén kártapasztalata. Ha kisebb, akkor pedig nagyobb súlya lesz μ-nek, azaz az egész portfólió tapasztalatának. Ez szintén ésszerűnek tűnik, hiszen minél többet van kockázatban egy szerződés, annál relevánsabb az általa gyűjtött tapasztalat.

4.3. Bühlmann-Straub modell

Amíg az előző példánkban Poisson eloszlású kárszám paraméterének Gamma eloszlást feltételeztünk, míg a Bühlmann-Straub modell esetén egy egész eloszlás feltételezése helyett csak az első két mometum tekintetében élünk feltételezésekkel. Ez előnyt jelenthet számunkra olyan esetekben, ahol nem ismerjük az egész eloszlást, csak a várható értékre és a szórásra van valamilyen sejtésünk illetve csak ezeket tudjuk megfigyelni.
Tartsuk meg az előző fejezet jelöléseit, legyen Y_{jt} a magyarázott változónk, ahol j a multifaktor szintjét jelölti, t pedig azt, hogy az adott szinten, melyik megfigyelésről van szó és legyen μ az egész portfóliónk tapasztalati várható értéke. Vezessünk be egy új jelölést, legyen u_j egy determinisztikus hatás (relativitás) a j. csoporton belül, amit nem ismerünk, viszont becsülhetünk GLM-mel abban az esetben, ha az adott szinten elegendő adatunk van. Mivel MLF-ről van szó, szinte biztosan lesz olyan szegmensünk, ahol nem lesz elegendő megfigyelésünk, hogy elfogadható becslést adjunk rá. Ebben az esetben azzal a feltételezéssel élünk, hogy a j. szint megfigyelései nem feltétlenül ugyanazt a relativitást hordozzák, viszont a relativitások ugyanazt az U_j eloszlást követik. Ekkor a következő feltételes várható értékre vagyunk kíváncsiak:

$$E(Y_{jt}|U_j) = \mu U_j. \tag{4.6}$$

Mivel az egész portfóliónk várható értéke μ, ezért érdemes feltennük, hogy $E(U_j) = 1$,

tehát a relativitások várható értéke 1 és e körül szóródnak, illetve kényelmi szempontok miatt, érdemes $V_j := \mu U_j$-t vizsgálnunk U_j helyett. Így:

$$E(Y_{jt}|V_j) = V_j. \tag{4.7}$$

Amikor GLM-mel dolgozunk gyakran feltételezzük, hogy az adataink Tweedie modellből származó eloszlást követnek (pl normális, Poisson, Gamma), ekkor pedig (2.5)-ös egyenlet és 2.7. lemma alapján igaz a következő összefüggés:

$$D^2(Y_{jt}|V_j) = \frac{\phi V_j^p}{\omega_{jt}} \tag{4.8}$$

ahol ϕ a skála paraméter. Tegyük fel, hogy V_j minden j-re azonos eloszlású, ekkor bevezethetjük a j-től független jelölést: $\sigma^2 := \phi E(V_j^p)$, így pedig (4.8) alapján:

$$E[D^2(Y_{jt}|V_j)] = \frac{\sigma^2}{\omega_{jt}} \tag{4.9}$$

Szedjük össze, milyen feltételezésekkel éltünk ebben az esetben:

- Az (Y_{jt}, V_j) vektorok függetlenek egymástól.
- A V_j-k azonos eloszlásúak és teljesül, hogy $E(V_j) = \mu > 0$, illetve $D^2(V_j) = \tau^2$, valamilyen $\tau > 0$-ra.
- Minden j-re, $Y_{jt}|V_j$ függetlenek egymástól. A várható értékét a (4.7)-es pont adja meg, a szórásnégyzet pedig kielégíti a (4.9)-es egyenletet.

Nem lett volna szükséges Tweedie modellet feltételeznünk, igazából csak a (4.9)-es képlet általános felírásához nyújtott nekünk motivációs segítséget. Írjuk fel Y_{jt} szórásnégyzetét is:

$$D^2(Y_{jt}) = D^2[E(Y_{jt}|V_j)] + E[D^2(Y_{jt}|V_j)] = \tau^2 + \frac{\sigma^2}{\omega_{jt}} \tag{4.10}$$

Amikor vizsgáljuk az adatainkat, ahogy már korábban is említettük, két eset lehetséges. Amennyiben elegendő adat áll a rendelkezésünkre, az adott szinten a magyarázó változók egyértelmű hatást hordoznak és ebben az esetben V_j becsléséhez jól használható a súlyozott átlag:

$$\overline{Y}_{j.} = \frac{\sum_t \omega_{jt} Y_{jt}}{\sum_t \omega_{jt}} \tag{4.11}$$

Abban az esetben, ha nem minden szegmensben rendelkezünk elegendő megfigyeléssel, akkor ott használhatjuk μ-t illetve annak valamilyen becslését. Amikor pedig V-t akarjuk megbecsülni, akkor a megfigyelések összes szóba jöhető lineáris függvényei között keressük azt, ami négyzetes értelemben a legjobban közelíti V-t.

4.1. Tétel (*Bühlmann-Straub*)**.** *A megfigyelések lineáris függvényei közül (V) a $E[(\hat{V} - V)^2]$ várható értéket a*

$$\hat{V}_j = z_j \overline{Y}_{j.} + (1 - z_j)\mu, \tag{4.12}$$

függvény minimalizálja, amit $\hat{V}$ lineáris credibility becslésének hívunk, és tudjuk hogy $z_j = \frac{\omega_{j.}}{\omega_{j.} + \sigma^2/\tau^2}$.

Itt μ-t ismerjük már a korábbi adatainkból, vagy meg tudtuk becsülni vagy használhatjuk a súlyozott átlagot (4.2). A τ^2-t és a σ^2-t becsléssel kaphatjuk meg, amit a következő részben részletesebben bemutatok.
Ezek alapján az U_j véletlen hatás credibility becslését a következő módon kaphatjuk meg: $\hat{U}_j = z_j \overline{Y}_{j.}/\mu + (1 - z_j)$.

4.4. Paraméterek becslése

A credibility elmélet a sok u_j relativitás megbecslése helyett csak a μ, a σ^2(ami a csoportokon belüli variancia) és a τ^2 (ami pedig a csoportok közötti variancia egyfajta mértéke) becslésére korlátozza le a becsülendő paraméterek számát.
A becsléshez használhatjuk az egész portfóliónkat, ami azért könnyíti meg a helyzetünket, mert nem okoz problémát, ha vannak olyan csoportok, ahol kevés a megfigyelt adat. Vezessük be az n_j jelölést, ami a $j.$ szinten ismétlődő megfigyelések száma. Ekkor a σ_j^2-t a következő négyzetösszegekkel becsülhetjük meg torzítatlanul:

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_t \omega_{jt} (Y_{jt} - \overline{Y}_{j.})^2 \tag{4.13}$$

Pontosíthatjuk a becslésünket, illetve függetlenné tehetjük j-től, ha súlyozzuk a szabadságfokkal:

$$\hat{\sigma}^2 = \frac{\sum_j (n_j - 1)\hat{\sigma}_j^2}{\sum_j (n_j - 1)} \tag{4.14}$$

Hasonlóan $\tau^2 = D^2(V_j)$ becslése a $\sum_j \omega_{j.}(\overline{Y}_{j.} - \overline{Y}_{..})^2$ négyzetösszeg alapján történik. Amennyiben a becslésünket az alábbira módosítjuk a becslésünk továbbra is torzítatlan marad:

$$\hat{\tau}^2 = \frac{\sum_j \omega_{j.}(\overline{Y}_{j.} - \overline{Y}_{..})^2 - (J-1)\hat{\sigma}^2}{\omega_{..} - \sum_j \omega_{j.}^2/\omega_{..}}. \tag{4.15}$$

4.2. Tétel. *A Bühlmann-Straub modell esetén, a $\hat{\sigma}^2$ és a $\hat{\tau}^2$ becslések torzítatlanok, tehát:*

$$E(\hat{\sigma}^2) = \sigma^2, \qquad illetve \qquad E(\hat{\tau}^2) = \tau^2. \tag{4.16}$$

4.4.1. Numerikus példa a credibility becslésre

Most, hogy már minden tudás a kezünkben van, szeretném egy fiktív példán keresztül bemutatni, hogyan tud a biztosító az elmúlt év adataira, a saját megfigyeléseire és az egész biztosítási piac adataira támaszkodni a számolásai során. Vizsgáljunk egy csoportos egészségbiztosítást négy különböző cég esetében, ahol rendelkezésünkre állnak az elmúlt nyolc év adatai. Szeretnénk a következő évre, cégenként külön-külön meghatározni a nettó kárszükségletet egy-egy munkavállalóra vetítve. Rendelkezésünkre állnak mind a négy vállalat esetében az elmúlt nyolc év adatai. Ismerjük, hogy melyik cégnél, melyik évben mekkora volt a kárkifizetés (eFt) és tudjuk még, hogy az adott évben átlagosan mennyi volt az aktív dolgozók száma (fő) a vállalatnál. Az adataink a következőek:

Év	*Cég(1)*	*Cég(2)*	*Cég(3)*	*Cég(4)*
1	526 (42)	97 (19)	0 (7)	281 (22)
2	441 (46)	104 (21)	402 (6)	289 (22)
3	530 (52)	167 (25)	1057 (11)	431 (19)
4	487 (58)	103 (22)	317 (7)	199 (21)
5	481 (58)		0 (9)	657 (14)
6	499 (55)		821 (5)	181 (10)
7	412 (51)		0 (4)	521 (11)
8	590 (51)			201 (9)
$\overline{Y}_{j.}(\omega_{j.})$	495,75 (413)	117,75 (87)	371 (49)	345 (128)

A cellákban az adott évben, az adott cégnél megállapított kárszükséglet található, illetve zárójelben az adott évben ott dolgozók száma osztva tízzel. Például a második számú cégnél a 3. évben 167 egység volt a nettó kárszükséglet és átlagosan 250-en dolgoztak ott abban az évben. A táblázat utolsó oszlopában kiszámoljuk $\overline{Y}_{j.}$ átlagot és az $\omega_{j.}$ szumma értékeket. Számoljuk ki $\tilde{\mu} = \overline{Y}_{..}$ értéket, azaz az egész piac átlagát.

$$\tilde{\mu} = \overline{Y}_{..} = \sum_{j=1}^{4} \sum_{t=1}^{8} \frac{Y_{jt}\omega_{jt}}{\omega_{..}} \approx 412 \tag{4.17}$$

Most meghatározzuk a (4.14) és a (4.15) képletek segítségével σ^2 és τ^2 értékeket. Először a (4.13) egyenlet segítségével kiszámoljuk $\tilde{\sigma}_j^2$ értékét, ahol n_j jelöli az adott cégen belüli megfgyelések számát, azaz jelen esetben azon évek számát, amelyekhez tartozik megfigyelés. A kiszámolt értékek: $\sigma^2 = 752,32^2$ és $\tau^2 = 155,93^2$. Ezek segítségével és a Bühlmann-Straub tétel segítségével már mindent ki tudunk számolni. Készítsünk a könnyebb átláthatóság kedvéért egy összefoglaló táblázatot:

Változó	*Cég(1)*	*Cég(2)*	*Cég(3)*	*Cég(4)*
$\omega_{j.}$	413	87	49	128
$\overline{Y}_{j.}$	497,75	117,75	371	345
z_j	0,95	0,79	0,68	0,85
$\hat{V}_j$	491,3	179,9	384,2	355,3
$\hat{U}_j$	1,192	0,437	0,933	0,862

Érdemes megvizsgálnunk a táblázatból a z_j értékeket, ami azt mutatja meg, hogy az adott cégek esetében, mekkora részben támaszkodott a modell a vállalati adatokra és milyen részben az egész piac átlagára. Jól látható, ahol több megfigyelésünk volt ott nagyobb a z_j érték, ahol kevesebb ott kisebb ez az érték. Az is jól látható, hogy az egész piac átlaga 412 egység, viszont az egyes cégek várható nettó kárszükséglete ettől jelentős mértékben eltér.

4.5. A Klasszikus Bühlmann-Straub modell és a GLM együttműködésben

A klasszikus Bühlmann-Straub modell a kárszükséglet meghatározásához akkor is alkalmazható , amikor több multifaktort kell egyidejűleg kezelnünk. Egy példán keresztül megmutatjuk, hogyan kombinálható az általánosított lineáris modell és a credibility elmélet a gyakorlatban.

Vizsgáljunk egy lakásbiztosítást, ahol a következő megfigyeléseket vonjuk be a vizsgálatunkba: a lakás földrajzi elhelyezkedése az irányítószám alapján (ez egy multifaktor), a lakás alapterülete (csoportokba osztva, hogy ne folytonos, hanem kategorikus változó legyen) és a lakás típusa (panel, családi stb. ami szintén kategorikus változó is). Ekkor utóbbi kettőt GLM-mel tudjuk becsülni, míg az első változó hatásának becsléséhez a credibility elméletre is szükségünk van. Ekkor (4.6) alapján a következőt írhatjuk fel:

$$E(Y_{ijt}|U_j) = \mu\gamma_1^i\gamma_2^i U_j. \tag{4.18}$$

Itt i-vel indexeljük azon változók szintjeinek sorszámozását, amelyeket standard GLM módszerrel tudunk becsülni, j továbbra is a multifaktor szintjét, t pedig a multifaktor szintjén belül a megfigyelés sorszámát jelöli és függ a szinttől is. Továbbá μ jelöli a bázis kárszükségletet, γ_1^i hordozza a ház típusa szerinti, míg γ_2^i a ház területe szerinti információt, U_j pedig a multifaktor véletlen hatását tartalmazza az adott szinten. Hogy meghatározhassuk a biztosítás nettó díját, meg kell becsülnünk GLM-mel a γ_1^i-et és a γ_2^i-öt, illetve U_j-t a credibility becsléssel.

Általánosítsuk annyiban a modellünket, hogy ne csak két ordinális magyarázó változót vegyünk be a modellbe, hanem R különböző faktort, amelyek mind standard GLM technikával becsülhetőek. Ebben az esetben a multiplikatív modellünk a következő alakot ölti:

$$E(Y_{ijt}|U_j) = \mu\gamma_1^i\gamma_2^i \dots \gamma_R^i U_j. \tag{4.19}$$

A bázis kárszükséglet, egy olyan szerződés kárszükséglete, melynek minden tulajdonsága az adott faktor bázisszintjének felel meg, tehát ahol $\gamma_r^i = 1$; $r = 1, \dots, R$, illetve ahogy korábban most is tegyük fel, hogy $E(U_j) = 1$ teljesüljön. Egyszerűsítsük a modellünket a következő jelölések bevezetésével, ahol tegyük fel, hogy ezek a standard relatívitások már ismertek:

$$\gamma_i = \gamma_1^i\gamma_2^i \dots \gamma_R^i \tag{4.20}$$

mivel ez csak lényegében i-től függ i kódolja mind az R faktor szintjét, illetve legyen $V_j := \mu U_j$ ismét, így pedig:

$$E(Y_{ijt}|V_j) = \gamma_i V_j \tag{4.21}$$

alakra egyszerűsödött a modellünk. Ahogy már korábban is tettük, éljünk azzal a motivációs feltételezéssel hogy $Y_{ijt}|V_j$ Tweedie modellhez tartozó valamely eloszlást követi, ekkor pedig:

$$D^2(Y_{ijt}|V_j) = \frac{\phi(\gamma_i V_j)^p}{\omega_{ijt}} \tag{4.22}$$

Ismét legyen $\sigma^2 := \phi E(V_j^p)$, így pedig:

$$E[D^2(Y_{ijt}|V_j)] = \frac{\gamma_i^p \sigma^2}{\omega_{ijt}} \tag{4.23}$$

fog teljesülni. Kibővítjük a 4.3.-as fejezetben tett feltételezéseinket:

- Az (Y_{ijt}, V_j) vektorok függetlenek egymástól.
- A V_j-k azonos eloszlásúak és teljesül, hogy $E(V_j) = \mu > 0$, illetve $D^2(V_j) = \tau^2$, valamilyen $\tau > 0$-ra.
- Minden j-re, $Y_{ijt}|V_j$ függetlenek egymástól. A várható értékét a (4.21)-es pont adja meg, a szórásnégyzet pedig kielégíti a (4.23)-es egyenletet.

Új jelöléseket vezetünk be, hogy visszavezessük a korábbiakra az eredményeinket.

$$\tilde{Y}_{ijt} = \frac{Y_{ijt}}{\gamma_i}, \qquad \tilde{\omega}_{ijt} = \omega_{ijt}\gamma_i^{2-p} \tag{4.24}$$

Így pedig fennállnak a következők:

$$E(\tilde{Y}_{ijt}|V_j) = V_j, \qquad E[D^2(\tilde{Y_{ijt}}|V_j)] = \frac{\sigma^2}{\tilde{\omega}_{ijt}} \tag{4.25}$$

Ezekkel a jelölésekkel kielégítjük a klasszikus Bühlmann-Straub tétel feltételeit, így jelen esetben akövetkező állítást kapjuk:

4.3. Tétel (*Klasszikus Bühlmann-Straub*)**.** *A lineáris $\hat{V}$ függvények között a $E[(\hat{V} - V)^2]$ várható értéket a*

$$\hat{V}_j = \tilde{z}_j \overline{\tilde{Y}}_{.j.} + (1 - \tilde{z}_j)\mu, \qquad ahol \qquad \tilde{z}_j = \frac{\tilde{\omega}_{.j.}}{\tilde{\omega}_{.j.} + \sigma^2/\tau^2}. \tag{4.26}$$

függvény minimalizálja.

Ekkor a korábbiakkal analóg módon fennáll:

$$\hat{U}_j = \tilde{z}_j \frac{\overline{\tilde{Y}}_{.j.}}{\mu} + (1 - \tilde{z}_j) \tag{4.27}$$

Amikor Y_{ijt} a kárgyakoriságot jelöli, ω_{ijt} pedig a kitettséget, akkor a kettő szorzata a kárszámot adja, így ebben az esetben a GLM-et Poisson eloszlás feltételezésével illesztjük, és U_j becslése:

$$\frac{\overline{\tilde{Y}}_{.j.}}{\mu} = \frac{\sum_{i,t} \omega_{ijt} Y_{ijt}}{\sum_{i,t} \omega_{ijt} \mu \gamma_i} \tag{4.28}$$

lesz, ami egy természetes becslése, hiszen a károk száma a j-edik csoportban osztva a várható kárszámmal ugyanabban a csoportban.

Amikor Y_{ijt} a kárnagyságot jelöli, ω_{ijt} pedig a kárszámot, akkor a GLM-et Gamma eloszlás feltételezésével illesztjük, így U_j becslése a következő lesz:

$$\frac{\overline{\tilde{Y}}_{.j.}}{\mu} = \frac{\sum_{i,t} \omega_{ijt} Y_{ijt}/(\mu\gamma_i)}{\sum_{i,t} \omega_{ijt}} \tag{4.29}$$

lesz, ami szintén egy logikus becslés, hiszen a j csoport átlagkára osztva a csoport várható átlagkárával és ez kisúlyozva a kárdarabszámokkal.

A variancia paraméterek becslése ($\hat{\sigma}^2$ és $\hat{\tau}^2$), az előző fejezethez hasonló módon a következő torzítatlan becslésekkel számolható:

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{it} \tilde{\omega}_{ijt} (\tilde{Y}_{ijt} - \overline{\tilde{Y}}_{.j.})^2, \tag{4.30}$$

$$\hat{\tau}^2 = \frac{\sum_j \tilde{\omega}_{.j.} (\overline{\tilde{Y}}_{.j.} - \overline{\tilde{Y}}_{...})^2 - (J-1)\hat{\sigma}^2}{\tilde{\omega}_{...} - \sum_j \tilde{\omega}_{.j.}^2 / \tilde{\omega}_{...}}. \tag{4.31}$$

4.6. A backfitting algortimus

Az előző fejezet jelöléseit felhasználva, bemutatjuk milyen algoritmus használható a gyakorlatban. Amennyiben ismerjük a γ_i értékeket és μ-t, akkor ezekből becsülni tudjuk $\hat{U}_j$ értékét a (4.27)-es képlet segítségével. Ha viszont U_j-t vagy annak valamilyen becslését ($\hat{U}_j$) ismerjük, akkor becsülhetjük ennek segítségével μ-t illetve a $\gamma_1^i, \ldots, \gamma_R^i$-ek értékét GLM-mel, amiből pedig ki tudjuk számolni a γ_i-t is. Ekkor úgy kezeljük az $\hat{U}_j$-t, mint ismert hatást *(offset)* és beépítjük a GLM-be. Szóval ha U_j-t ismerjük, akkor becsülhetjük μ-t és γ_i-t, illetve fordítva is igaz ez. Ez pedig arra ösztönöz minket, hogy

egy iteratív algoritmust használjunk. Ezt az eljárást nevezik *(backfitting algoritmus)*-nak, ami iteratív módon, szimultán határozza meg a szükséges paramétereket - a faktorokat GLM segítségével, a multifaktorokat pedig credibility becsléssel -, a következő lépések iterálásával:

1. lépés: Először is legyen minden $\hat{U}_j = 1$.

2. lépés: Alkalmazzunk GLM-et a következők szerint: válasszunk valamilyen Tweedie modellt (általában Gamma vagy Poisson), a link függvényünk legyen a logaritmus, a (multiplikatív) modellbe pedig építsük be offsetként $log(\hat{U}_j)$-t. (Ekkor megkapjuk $\hat{\mu}$-t és $\hat{\gamma}_1^i, \ldots, \hat{\gamma}_R^i$ becsléseket.)

3. lépés: Számoljuk ki $\hat{\sigma}^2$ és $\hat{\tau}^2$ értékét (4.30), illetve (4.31) alapján.

4. lépés: Számoljuk ki $\hat{U}_j$ értékét (4.27) alapján.

5. lépés: Térjünk vissza a második lépésre, és folytassuk az új $\hat{U}_j$-vel az algoritmust.

Alkalmazzuk az iterációt, amíg konvergál. Ez gyakran nagy ismétlés számot igényel a pontos közelítéshez, általában 100 lépés szükséges egy nagyjából 10^{-4}-es pontosság eléréséhez. Természetesen vannak esetek, amikor a konvergencia sokkal gyorsabb.

A gyakorlatban gyakran előfordul, hogy egynél több multifaktorunk van. Például egy kötelező gépjármű felelősségbiztosítása esetén, ha osztályozzuk a gépjármű típusa, és a lakóhely szerint is a biztosítottakat, akkor máris két különböző MLF-ünk van. Ebben az esetben sem kell kétségbe esni, az algoritmus ugyanúgy használható, ahogy egy darab multifaktor esetében. Annyi változik, hogy mind a kettőt offsetként beépíthetjük a modellünkbe, csak arra kell figyelnünk, amíg az egyik szerint futtatjuk az algoritmusunkat, addig a másikat kezeljük úgy, mint egy standard változót (ami nem multifaktor és ismerjük már a hatását). Az említett példában először a gépjármű típusa szerint szokták futtatni az algoritmust, majd utána a területi elhelyezkedés szerint.

4.7. Hierarchikus credibility modellek

Teszek egy kis kitérőt és bemutatom a hierarchikus credibility modellek elméletét. A 4.1.-es fejezet témája a multifaktorok voltak, amelyek olyan változók, amik a kezelhetőnél több szinttel rendelkeznek. A gyakorlatban nem csak ilyenek fordulnak elő, hanem egymásba ágyazott multifaktorok is, azaz olyan változók, amelyeknek minden egyes szintjén egy újabb ilyen található, ami így első hallásra nagyon ijesztőnek hathat a kezelhetőség szempontjából.

Tipikusan ilyen egymásba ágyazott multifaktorok a gépjárművek és azoknak a márkája, a modellje majd a típusa. Például Skoda lehet a gépjármű márkája, a modellje Fabia, a típusa pedig az adott modell egyedi verziója. Erre a rendszerre épül például a eurotax, ami használt gépjárművek piaci értékének meghatározásával foglalkozik, ami jó alapot nyújt a biztosítóknak is a CASCO biztosításnál a gépjárművek értékeléséhez.

Egy másik, a gyakorlatban gyakran előforduló példa az egymásba ágyazott multifaktorokra a területi szegmentáció. Ebben az esetben a legfelsőbb szintet a megyék jelentik, amik kistérségekre vannak felosztva, ahol pedig települések találhatóak, sőt gyakran a településeken belül még eltérő irányítószámok is lehetnek.

Nézzük, hogy működnek a hierarchikus credibility modellek. Hagyományos credibility modellhez hasonlóan működnek, azzal a különbséggel, hogy egyszerre adnak becslést minden hierarchiai szinten levő paraméterre majd az így kapott eredményeket beépíthetjük a modellünkbe, mint ismert hatásokat.

5. fejezet

Számolások R-ben

5.1. Az adatok

Szeretném diplomamunkám zárásaként az általánosított lineáris modell használatát bemutatni egy valós megfigyeléseket tartalmazó adathalmazon az R program segítségével. A [3]-as szakirodalom 2. fejezetének végén található esettanulmány publikusan elérhető adatait használom, amelyek a `http://www2.math.su.se/~esbj/GLMbook/case.html` oldalon *mccase.txt* néven elérhetőek mindenki számára.
Az itt található adatok a korábbi *Wasa* svéd biztosító 1994-1998 közötti CASCO állományát tartalmazzák. A letölthető fájl a következő megfigyeléseket tartalmazza: a tulajdonos kora, neme és lakóhelye, ami már 7 különböző zónára van felosztva, a gépjármű kora, a vezető bónusz besorolása, ami szintén 7 szinttel rendelkezik, a kockázatban töltött idő, a kárszám és a kárnagyság, illetve egy sajátos mérőszám, amit a motor teljesítményéből és súlyából számolnak ki, ami alapján 7 különböző osztályt hoznak létre.

- **Nem:** A vezető neme. Két szinttel rendelkező faktor.
- **Kor:** A vezető kora. Egészértékű változó.
- **Zóna:** Területi besorolás, hét szinttel rendelkező faktor.
- **Osztály:** Hét szinttel rendelkező faktor, amit a biztosító a következőképpen számol. A motor teljesítménye kw-ban, az szorozva százzal, majd osztva a gépjármű

tömegével plusz 75 kg-mal majd egészekre kerekítve. (A 75 kg egy átlagos vezető súlya.)

- **Bónusz:** A vezető bónusz besorolása. Hét szinttel rendelkező faktor.
- **Kitettség:** Kockázatban töltött idő. Numerikus változó.
- **Kárszám:** Károk száma. Egészértékű változó.
- **Kárnagyság:** Károk nagysága. Folytonos változó.

5.2. Az R program

Szerencsés helyzetben vagyunk, hiszen a diplomamunkám során bemutatott módszereket nem kell kézzel számolgatnunk, manapság már rengeteg olyan statisztikai program létezik, amik megteszik ezt helyettünk. Ilyen többek között az R program is, amit jómagam is használtam.
A letöltött fájlban az adatok már osztályokba sorolva találhatóak - például zóna alapján, - így nekem ezzel már nem kellett foglalkoznom, pedig az R-ben található *tree* paranccsal elvégezhetnénk az adatok csoportosítását. Ami számunkra érdekes lehet, az kiválasztani a szignifikáns változókat, amelyeket beveszünk a modellünkbe. A különböző modelleket az R a 3.2.-es fejezetben tárgyalt AIC mutatók alapján hasonlítja össze. Egy nagyon hasznos parancs a *step*, amely a különböző magyarázó változókat veszi sorra, majd hagyja ki azokat egyesével a modellünkből, így vizsgálva az AIC mutatókat, megkönnyítve ezzel a munkánkat. Emellett, amikor az adatainkra GLM-et akarunk illeszteni, azt is kiválaszthatjuk, hogy melyik magyarázó változóink között kívánjuk a 3.3.-as fejezetben tárgyalt hatásokat is kezelni.

5.3. GLM illesztése R-ben

Amikor a biztosító GLM-et akar illeszteni az adataira a következőkre kell figyelnie. Mi legyen a függő változó (kárszám, kárnagyság, kárgyakoriság, kárszükséglet), amit

modellezni szeretne. Van-e olyan változó, amit offsetként beépíthet a modellbe? Milyen eloszlást (Poisson, Gamma, normális, inverz Gauss stb.) feltételez az adatoknak, és milyen link függvényt választ ehhez, illetve melyek azok a változók, amelyek között hatást feltételez? Ha kiválasztja a függő változót, amit modellezni szeretne, ahhoz általában már tudjuk, hogy milyen eloszlás fog jól illeszkedni (pl. kárszám $\sim$ Poisson, kárgyakoriság $\sim$ Gamma). Offsetként érdemes a modellbe beépíteni kárszám esetén a kitettséget (illetve annak a logaritmusát), míg átlagkár esetén a kárszámot. Mivel az eloszlásokhoz tartozik kanonikus link függvény érdemes azt választanunk, de abban az esetben, ha különböző modelleket akarunk összehasonlítani, akkor lehet, hogy a különböző eloszlások feltételezése miatt, különböző link függvényeink lesznek, de érdemes lehet nekünk azonos link függvénnyel számolni, mert akkor a kapott paramétereink azonos skálán lesznek értelmezve, így összehasonlíthatjuk azokat. A változók közötti hatások kiszűréséhez pedig kisérletezésre van szükség.

Én több modellt illesztettem, különböző eloszlás feltételezésével, különböző függő változókra, úgy hogy a modellbe bevett változókat is variáltam, majd a kapott AIC mutatókat hasonlítottam össze. Először legyen a függő változónk a kárszám, az offset pedig log(kitettség). A modellben Poisson eloszlást feltételeztem, a link függvény pedig legyen a logaritmus. A modell építésénél érdemes egyesével bevenni a különböző magyarázó változókat, és vizsgálni az AIC mutatóikat, hogy lássuk, tudtunk-e javítani a modellünkön. Az alábbi táblázatok tartalmazza a lényeges információkat:

Magyarázó változó	**Kor**	**Nem**	**Zóna**	**Osztály**	**Gépkor**	**Bónusz**	**AIC**
Modellben?	+	-	-	-	-	-	7.339,7
Modellben?	+	+	-	-	-	-	7.319,8
Modellben?	+	+	+	-	-	-	7.170,9
Modellben?	+	+	+	+	-	-	7.032,3
Modellben?	+	+	+	+	+	-	6.966,3
Modellben?	+	+	+	+	+	+	6.947,2

5.1. táblázat. GLM(Kárszám $\sim$ family=Poisson(link=log), offet=log(kitettség))

A modellt vizsgáltam offset nélkül és offsettel is. Amikor a kitettséget, mint ismert hatást kezeltem, akkor kisebb AIC mutatókat kaptam, így ezt megtartottam a modellben. A publikusan elérhető adatokban a bónusz egy hétszintű faktorként jelent meg, így magyarázó változóként tudtam beépíteni a modellbe. A gyakorlatban gyakran ezt is offsetként építik be, ezzel én is próbálkoztam, de nem tudtam olyan szorzókat generálni, amivel jobb illeszkedést kaptam volna. Ha ez sikerült volna, akkor kevesebb paraméter becslésére lett volna szükség, hiszen a bónuszt, mint ismert hatást beépítettük volna a modellünkbe, így azt már nem kellett volna tovább becsülnünk. A táblázatból jól látható, hogy minden egyes változóval, amit a modellünkhöz hozzávettünk, javítottuk az illeszkedést. Egyelőre viszont nem vettük figyelembe a változók közötti rejtett hatásokat, teszteljük ezt is, hátha tovább tudjuk javítani a modellünket. Tartsuk meg az eddig bevett változókat, és vizsgáljuk tovább a modellünket úgy, hogy figyelünk a változók közötti hatásokra. Ezt nem részletezném táblázatos formában, de a kapott eredményeim alapján a kor, a nem és a bónusz besorolás között van interakció, így azokat érdemes bevonni a modellünkbe, így az AIC mutatónk 6.924,6 lett. Ugyanezt a modellt lefuttattam azzal az eltéréssel, hogy egyszer kvázi Poisson eloszlást, egyszer pedig negatív binomiális eloszlást feltételeztem, ezek azonban rosszabb illeszkedést mutattak, így ezt most szintén nem részletezném.

Vizsgáljuk meg azt az esetet, amikor a modellünket a kárnagyság/kárszámra próbáljuk meg illeszteni. Ezt is Gamma eloszlás, inverz Gauss eloszlás és normális eloszlás feltételeze eseteire vizsgáltam, log link függvénnyel, a modellbe offsetet nem építettem be. Ebben az esetben is a Gamma eloszlás mutatta a legjobb értékeket, így csak azt részletezném:

Magyarázó változó	**Kor**	**Nem**	**Zóna**	**Osztály**	**Gépkor**	**Bónusz**	**AIC**
Modellben?	+	-	-	-	-	-	14.583
Modellben?	+	+	-	-	-	-	14.582
Modellben?	+	+	+	-	-	-	14.567
Modellben?	+	+	+	+	-	-	14.463
Modellben?	+	+	+	+	+	-	14.446
Modellben?	+	+	+	+	+	+	14.443

5.2. táblázat. GLM(Kárnagyság/kárszám $\sim$ family=Gamma(link=log)

Láthatjuk, hogy ebben az esetben ismét érdemes volt minden változót bevenni a modellünkbe. Ebben az esetben is vizsgáltam interakciót, meglepő módon az eddigiekkel szemben itt nem tudtam jelentős csökkenést elérni az AIC mutatóban különböző interakciók feltételezésével a modellben. A táblázatokban nem részleteztem az iterációk lépésszámát, ami az egyszerűbb (pár változós) modellek esetén 7 és 9 között volt, míg a bonyolultabbak esetén 12 és 14 között, illetve az inverz Gauss eloszlás esetén nem konvergált az illesztő algoritmus.

6. fejezet

Összefoglalás

Kisebb hiányérzete lehet a diplomamunkám olvasásakor a tisztelt Olvasónak, amin nem csodálkozom hiszen nekem is az van. Amikor kiválasztottam az általam legérdekesebbnek vélt témát, az általánosított lineáris modellekről, nem gondoltam volna, hogy ekkora fába vágom a fejszémet. Az anyag amellett, hogy nagyon érdekes, roppant szerteágazó is, sajnos a dolgozatom terjedelmi korlátja miatt, nem jutott mindenre kellő figyelem.

Amikor nekifogtam a diplomamunkám írásának arra gondoltam, hogy nem csak a száraz matematikai és statisztikai hátterét fogom vizsgálni, hanem könnyen érthető példákon keresztül megpróbálom megértetni a modell működését is, hogy a kevésbé szakavatott Olvasó is érdekesnek találhassa a munkámat. Valójában a dolgozatom minden egyes fejezetéből nyugodtan lehetett volna egy különálló, komplex diplomamunkát írni, így viszont, hogy szerettem volna az egész modellt bemutatni, a kialakulásától kezdve, különböző jól érhető példákon át, egészen a számítástechnikai alkalmazásokig, talán egyik részt sem tudtam kellő mélységekig feldolgozni.

Akinek sikerült felkeltenem az érdeklődését a téma iránt, (esetleg tényleg hiányérzete van), azoknak bátran ajánlhatom az irodalomjegyzékben található első három könyvet tanulmányozásra, amelyek különböző szemszögből, különböző matematikai bonyolultsággal közelítik meg az általánosított lineáris modelleket, így mindenki megtalálhatja a számára leginkább megfelelőt.

Irodalomjegyzék

[1] DUNCAN ANDERSON, SHOLOM FELDBLUM, CLAUDINE MODLIN,DORIS SCHIRMACHER, ERNESTO SCHIRMACHER, NEEZA THANDI

A Practitioner's Guide to Generalized Linear Models – A foundation for theory, interpretation and application

Wattson Wyatt Worldwide, New York, February 2007.

[2] PIET DE JONG, GILLIAN Z. HELLER

Generalized Linear Models for Insurance Data

Cambridge University Press, NewYork, 2008.

[3] ESBJÖRN OHLSSON, BJÖRN JOHANSSON

Non-Life Insurance Pricing with Generalized Linear Models

Springer, 2010.

[4] MICHAEL J. CRAWLEY

The R Book

John Wiley & Sons Ltd., London, 2007.

[5] ARATÓ MIKLÓS

Általános biztosításmatematika

ELTE Eötvös kiadó, 1997.

[6] http://en.wikipedia.org