Biztosítói káradatok matematikai modellezése

Szakdolgozat

Készítette: Sebők Tamás

Matematika B.Sc., Matematikai elemző szakirány

Témavezető: Zempléni András, egyetemi docens

Valószínűségelméleti és Statisztika Tanszék

Eötvös Loránd Tudományegyetem
Természettudományi Kar

2015

Tartalomjegyzék

1. fejezet

Bevezetés

1.1. Motiváció

A témaválasztásnál az egyik legfontosabb szempont volt, hogy a későbbi munkám során fel tudjam használni a feltárt eredményeket, következtetéseket, és iránymutatást kapjak a káradatok jellemzőiről. Korábban dolgoztam több biztosítótársaságnál is, jelenleg pedig banki területen tevékenykedem, így számomra testhezálló feladat e téma kidolgozása. E miatt döntöttem a - biztosítói káradatok - téma választása mellett.

1.2. Feladat ismertetése

A biztosítótársaság szemszögéből nézve fontos a káradatok elemzése az üzletmenetre illetve stratégiára kiható információk leszűrésére. Továbbá meghatározó a káradatok elemzése tanulmányozása, ehhez szükséges egy megfelelő adathalmaz amiből a következtetések már érdemben levonhatók. A kapott 30000 sorból, két eredményváltozóból és 25 magyarázóváltozóból álló adathalmaz már ilyen. A rekordok egy-egy ügyfelet reprezentálnak. Az első eredményváltozó a kárszám, a második a kárnagyság $[0 : 1]$ intervallumskálán. A 25 magyarázóváltozó pedig, az ügyfélhez kapcsolódó egy-egy jellemzőt ír le. Tehát van egy mintánk amiből becsléseket, hipotéziseket állíthatunk fel annak igazolására, hogy a magyarázóváltozó és az eredményváltozók között mekkora a függőség. Mintákból még az is megállapítható, hogy milyen trendek figyelhetők meg. A káradatok elemzésének azért is nagy a jelentősége, mert egy adott jellemzővel rendelkező ügyfélhez egy konkrét kárszám, illetve kárnagyság-eloszlás rendelhető, ami befolyásolja a biztosító kiadásait.

1.3. Alkalmazott programok, eljárások

Az egyszerűség kedvéért a szakdolgozat elkészítéséhez Ubuntu Linux operációs rendszert gedit szövegszerkesztőt és terminál ablakot használtam.

Az alapfeltételeken túlmenően, szükség van egy statisztikai és egy szedő programra is. Elemzések elkészítéséhez az R-et, szedőként pedig a LaTeX-et használtam. Ezen programok széles körben ismertek, ezért úgy gondolom a bemutatásuk nem szükséges. A publikálás során azonban felmerül egy olyan kérdés amire érdemes kitérni. Az eredmények vagyis az R program outputjai a fordítás helyén vagyis a terminálablakban jelennek meg, amit végül is egy dokumentumban szeretnénk látni.

1.3.1. Sweave parancs

Eredmények elkészítése az R programban történik, itt megjelenik rögtön az output. Az inputokat érdemes egy külön fájlba kimenteni a saveHistory paranccsal. Egy dokumentumban, hogy ne csak az input és output jelenjen meg, hanem a hozzá kapcsolódó értelmező szöveg is, ahhoz más parancs illetve program is szükséges. A dolgozatom szedéséhez a LaTeX alkalmazást használom, ezért szükségszerű egy olyan parancs ami az R kódokat futtatja a többit figyelmen kívül hagyja. Valamint elvárás, hogy az output egy TeX kiterjesztésű fájlként jelenjen meg. Erre a problémára megoldásként szolgál a Sweave parancs. Ez egy R-beli parancs, szintaxisa Sweave('fájlnév.rnw'), ami létrehozza a kívánt fájlt, természetesen meg kell neki adni, hogy mit fordítson és mit ne módosítson, amit megfelelő tageléssel érhetjük el. Ennek két változata van, egyik a szöveges másik a grafikus eredmények megjelenítésére szolgál. A fordítás előtt létre kell hozni egy speciális rnw kiterjesztésű fájlt, ami tartalmazza a TeX-es elemeket, valamint az R kódokat. Ezen túlmenően a tag-elések argumentumaiban finomra lehet hangolni a TeX-es outputot. Többek között arra is lehetőség van, hogy csak az R output jelenjen meg az input ne. Lehetőség van az ábrák címkézésére is, valamint az egyes R kódok megjelölésére, ezzel a megoldással az egész dolgozat dinamikussá válik.

2. fejezet

Választott módszerek

Ebben a fejezetben az elemzési módszerek elméletét fogom bemutatni, körbejárni. Mintáról van szó, ezért a feladatok a becslések, hipotézisek felállításáról, ellenőrzéséről fognak szólni. Egy-két elemi statisztikai megállapítást fogok tenni az elemzési részben (itt nem térek ki rá). Gondolok itt az adott változó milyen skálán mozog, milyen az eloszlása stb. A módszerek irányvonalat a függetlenségvizsgálat, a logisztikus regresszió, a dummy változók és a díjkalkuláció fogja adni.

2.1. Függetlenségvizsgálat

Kárszám, kárnagyság és a magyarázóváltozó között fontos megnézni a kapcsolatuk szorosságát, függetlenségét, függvényszerű kapcsolatát. Fontos tény, hogy csak mintával rendelkezünk ezért a sokaság teljes számbavétele nem lehetséges, csak mintából történő következtetés, aminek legfőbb eszköze a hipotézisvizsgálat. Függetlenségvizsgálat az egy hipotézisvizsgálat, ahol a nullhipotézis a függetlenséget jelenti, az alternatív hipotézis pedig ennek az ellentétét. Ebből következik, hogy az alternatív hipotézis elfogadásakor sztochasztikus vagy függvényszerű kapcsolatunk lehet.
Tehát a függetlenségvizsgálat során a:

$$H_0 : P_{ij} = P_{i\bullet} \cdot P_{\bullet j} \quad (i = 1,2,\ldots,s \text{ és } j = 1,2,\ldots,t)$$

nullhipotézist a

$H_1 : \exists$ olyan i és j amelyre $P_{ij} \neq P_{i\bullet} \cdot P_{\bullet j}$ (ha $i = 1,2,\ldots,s$ és $j = 1,2,\ldots,t$)

alternatív hipotézissel szemben teszteljük, ahol:
P_{ij} az első ismérv i -edik a második ismérv j -edik értékének együttes előfordulásának a valószínűsége a sokaságban.
$P_{i\bullet}$ és $P_{\bullet j}$ a peremeloszlás megfelelő valószínűségei.

Legyen: $v_{i\bullet}$ $(i = 1, \ldots, s)$ az első ismérv szerinti i-edik osztályhoz tartozó gyakoriság a mintánál és $v_{\bullet j}$ $(j = 1, \ldots, t)$ a második ismérv szerinti j-edik osztályhoz tartozó gyakoriság. Továbbá legyen $v_{ij} (i = 1, \ldots, s \; ; \; j = 1, \ldots, t)$ az első ismérv szerinti i-edik a második ismérv szerint a j-edik osztályhoz tartozó együttes gyakoriság.
Fontos megállapítás, hogy:

$M(v_{i\bullet}) = nP_{i\bullet}$, $M(v_{\bullet j}) = nP_{\bullet j}$, $M(v_{ij}) = nP_{ij}$,ahol $(i = 1, \ldots, s; \; j = 1, \ldots, t)$

M jelöli a várható értéket. Tehát a relatív gyakoriságokkal lehet becsülni a valószínűséget. Ha ismerjük a peremvalószínűségeket akkor tiszta, ha nem akkor becsléses függetlenségvizsgálatról tudunk beszélni. Jelen feladat során is kizárólag a mintára tudunk hagyatkozni, ezért a peremvalószínűségek nem ismertek, csak becsülni tudjuk a minta gyakoriságok alapján. Tehát most a becsléses esettel kell dolgoznunk. Ennek megfelelően a szabadságfok is változik.
Szükségünk van egy próbafüggvényre is amivel majd ellenőrizni tudjuk hipotézisünket, ami a mért és elvárt értékek közötti eltérések négyzetes összegéből indul ki. Ezt az ellenőrző vizsgálatot sztandardizálva végezzük el. Tiszta függetlenségvizsgálatnál a próbafüggvény:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(v_{ij} - n \cdot P_{i\bullet} \cdot P_{\bullet j})^2}{n \cdot P_{i\bullet} \cdot P_{\bullet j}}$$

$n \to \infty$ mellett H_0 esetén aszimptotikusan χ^2 eloszlású $st - 1$ szabadságfokkal
Becsléses illeszkedésvizsgálatnál:

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(v_{ij} - \frac{v_{i\bullet} \cdot v_{\bullet j}}{n})^2}{\frac{v_{i\bullet} \cdot v_{\bullet j}}{n}} = \sum_{i=1}^{s} \sum_{j=1}^{t} \frac{(v_{ij} - v_{ij}^{\star})^2}{v_{ij}^{\star}}$$

$n \to \infty$ mellett H_0 esetén aszimptotikusan χ^2 eloszlású $(s - 1)(t - 1)$ szabadságfokkal.
Érdekes kérdés lehet még az elfogadási tartomány. Kézi számításnál a táblázatbeli szignifikanciaszintekhez tartozó kritikus értékekhez tudunk viszonyítani, de a programok és így az R is pontosan megadja a p értéket.

2.1.1. Asszociációs merőszámok

A kapcsolat erősségének a kimutatására a χ^2 statisztika nem teljesen alkalmas, hiszen az érték nagyban függ az elemszámtól, szabadságfoktól. Célszerű lenne egy olyan mérőszám ami 0 és 1 közé szorítja a mutatót.
A Φ együttható:

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

Különböző mintanagyságok így már összehasonlíthatóvá vállnak.
Kontigencia együttható (Pearson-féle C):

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

További előnye az előbbihez képest, hogy értéke 0 és 1 között marad.
Cramer féle V együttható:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}}$$

ahol k az oszlopok és a sorok száma közül a kisebb.

2.2. Logisztikus regresszió

0 és 1 értékű változóknál a regressziószámítás közvetlenül nem lehetséges, hiszen ha két ismérvértékünk van, akkor értelmetlen lehet bármilyen egyenes illesztése. A magyarázó változó befolyásolja az eredmény bekövetkezésének a valószínűségét, ezért legyen a bekövetkezés valószínűsége a függő változó. Ebben az esetben az értéke 0 és 1 közé esik, ami nem túl szerencsés, regresszió érdekében jó lenne egy tágabb intervallum. Nézzük meg a két érték, kár illetve nincs kár bekövetkezésének esélyét, egymáshoz való arányát amit odds-nak nevezünk.

$$odds_x = \frac{P_x}{1 - P_x} \text{ahol } P_x = P(K = 1|x)$$

Ekkor a logisztikus regressziónál legyen az odds logaritmusa a magyarázó változók lineáris függvénye.

$$ln(odds_x) = \epsilon_0 + \beta_1 x_1 + \cdots + \beta_n x_m$$

így:

$$odds_x = e^{\epsilon_0 + \beta_1 x_1 + \cdots + \beta_n x_m} = e^{\beta x + \epsilon_0}$$

Ebből a valószínűség:

$$P_x = \frac{odds_x}{1 + odds_x}$$

Így megkapjuk az adott változó bekövetkezésének valószínűségét. Fontos, hogy a logisztikus regressziót két értékkel rendelkező eredményváltozóknál használjuk.

2.2.1. Dummy változók

A nominális skálán mért tulajdonságokat számokkal kódolnunk kell. A kódolás legegyszerűbb esete, hogy egy adott ismérvváltozathoz hozzárendelünk egy természetes számot. Egy területi ismérv hozzárendelésnél ez teljesen tetszőleges is lehet. Két ismérvváltozatnál triviálisan 0 és 1. Ebben az esetben a 0 jelentheti egy tulajdonság hiányát is az 1 pedig a meglétét. Bár ez nem teljesen törvényszerű. (Pl.: nemeknél) Más a kérdés abban az esetben, ha több ismérvváltozat van mint kettő. Ilyen esetekben a kódolás történhet természetes számokkal, azonban két ismérvváltozathoz hozzárendelt szám között semmilyen következtetés nem vonható le. Ez adatelemzésnél próblémát okozhat. Megoldás a dummy változó. Dummy változó jelentése: egy adott ismérv ismérvváltozatának a megléte. (Pl.: Veszprém megyei vagy sem, férfi-e vagy sem.) Ha az adott tulajdonsággal rendelkezik akkor legyen 1, minden más esetben 0. Így egy nominális skálán mért tulajdonság, amelynek n ismérvváltozata van átalakítható $n-1$ dummy változóra, ez elegendő. Ha az első átalakított változó 0 és 1 értéket vehet fel, a második átalakított változó ugyancsak 0 és 1 értéket vehet fel, és így tovább $n-1$-ig, akkor az utolsó tulajdonság kódolható azonosan nullával. Ezt a szakirodalom kontroll-csoportként is definiálja, ennek megválasztása alapulhat gazdasági megfontoláson, de adhoc jellegű is lehet.

2.3. Díjkalkuláció

A biztosítási díj definició szerint kockázat átvállalásáért a biztosító által felszámított ár, a biztosítási védelemért a biztosított által fizetett ellenérték. Másnéven bruttó díj. A bruttó díj több részből tevődik össze.Kockázati díj az a díj, amelyet a kockázatért kérünk el. A szűkebb értelmezés szerint beszélünk nettó kockázati díjról. A kockázati díjon felül a vállalkozási díj van. Ez a díj fedezi az adminisztrációs díjakat és a nyereséget. Az ezzel növelt rész a bruttó biztosítási díj, amelyet a szerződőnek meg kell fizetni. Jelen esetben a díjnak a nettó kockázati díja és a biztonsági pótléka, együttesen kockázati díja érdekel. Ez az alapja a díjszámításnak. Nézzük meg részletesebben. Nettó kockázati díj: Tekintsünk egy szerződőt melyet különböző kár érhet. Egy adott kárhoz hozzá lehet annak a bekövetkezésének a valószínűségét rendelni. A lehetséges károk és a hozzá tartozó valószínűségek lineáris kombinációja a szerződő kockázatának a várható értéke. A szerződőtől a biztosító ezt a kockázatot vállaja át. Ezek az ügyfelek úgynevezett homogén kockázati csoportot képeznek, mely csoportoknak a kockázati díja a csoport aggregált kockázatán alapszik. Az aggregált kockázat ugyancsak egy valószínűségi változó. Ezeknek az aggregált kockázatoknak a számítása a díjkalkulációs elvek segítségével történik.

2.3.1. Díjkalkulációs elvek

A várható érték elv alapján kalkulált aggregált kockázati díj a legegyszerűbbek közé tartozik. Képlettel:

$$\Pi_E(Z) = (1 + a) * E(Z)$$

Ahol a arányossági tényező, $E(Z)$ az aggregált káreloszlás várható értéke. Z minden díjelvnél az aggregált kockázatot jelenti. Előnye, hogy két változó szükséges hozzá. Hátránya az, hogy független a szórástól, nagy szórásnál nem javasolt a használata. A szórás elv használata, képlettel:

$$\Pi_D(Z) = E(Z) + b * D(Z)$$

A b az arányossági tényező $E(Z)$ és $D(Z)$ az aggregált káreloszlás várható értéke és szórása. Harmadikként megemlíteném a szórásnégyzet elvet:

$$\Pi_V(Z) = E(Z) + d * V(Z)$$

Ahol d ugyancsak arányossági tényező $V(Z)$ pedig az aggregált káreloszlás varianciája.

3. fejezet

Adatelemzések R-ben

3.1. Adatok áttekintése

Rendelkezésünkre álló biztosítói káradatok 2 eredményváltozóból és 25 magyarázóváltozóból állnak össze. Az első eredményváltozó a kárszám, a második a kárnagyság. A biztosítók az ügyfelek adatait és saját káradataikat bizalmasan kell, hogy kezeljék, ezért kellett egy olyan eljárás, – standardizálás – ami lehetővé tette, hogy ne ismerjék fel az adataikat. Ez az oka, hogy kicsi a kárnagyság. Nézzük meg, hogy az egyes kárszámokból és kárnagyságokból mennyi van. Kontigenciatábla segítségével csoportosítsuk, nem bekövetkezettre és bekövetkezettre, valamint kifizetett és nem fizetett károkra. Jelöljük $xd1$-vel a kárszámot, úgy hogy két értéke legyen. Bekövetkezett kár és nincs kár. Jelöljük $xd2$-vel a standardizált kárnagyságot csoportosítva, kifizetett kárra és nem kifizetett kárra.

```
   xd2
xd1     0     1
  0 26389     0
  1   176  3435
```

Látjuk, hogy az esetek több mint 80%-ban nem volt kár, valamivel több mint 10%-ban volt. Érdekes még, hogy 176 esetben ugyan volt kár, de a biztosító valamilyen oknál fogva nem fizetett.

A további elemzés érdekében csoportosítsuk a két eredményváltozót ($x1$ a kárszám, $x2$ a kárnagyság)

$$x1 = \begin{cases} 3 & \text{ha } x1 > 3 \\ x1 & \text{egyébként} \end{cases}$$

$$x2 = \begin{cases} 0 & 0 \leq x2 < 0.1 \\ 0.1 & 0.1 \leq x2 < 0.2 \\ 0.2 & 0.2 \leq 2 < 0.3 \\ 0.3 & 0.3 \leq x2 \end{cases}$$

Az eredmény:

```
   x2
x1      0   0.1   0.2   0.3
  0 26389     0     0     0
  1   171  2548    29    19
  2     5   570    12     4
  3     0   240    10     3
```

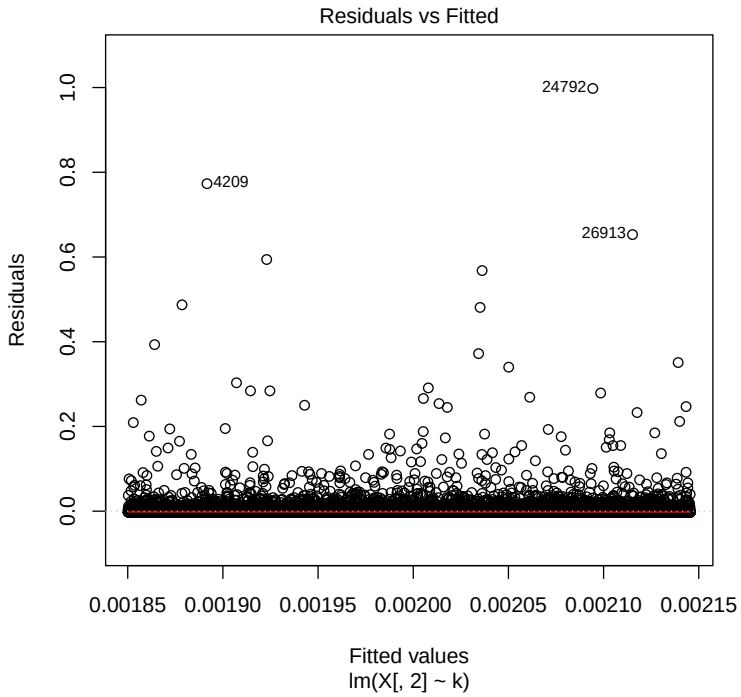
A kárszámok eloszlását érdemes egy táblázatban összefoglalni:

```
    0     1     2     3     4     5     6     7    11
26389  2767   591   164    62    16     6     4     1
```

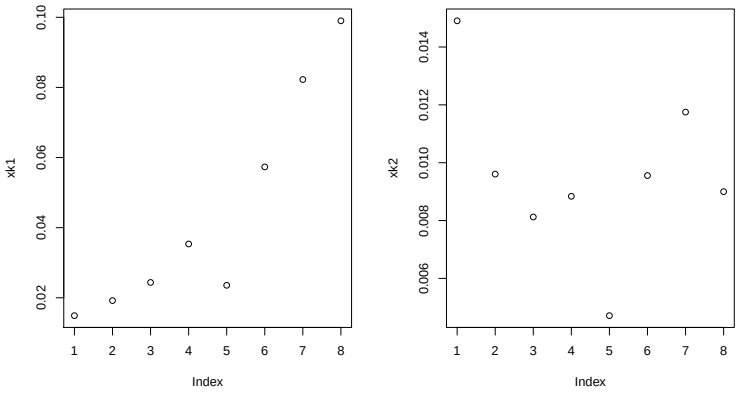
Érdekes kérdés lehet még, hogy a kárnagyságok és az indexük között van-e valamilyen kapcsolat. Ha esetleg van akkor az azt jelenti, hogy az adatok egy hosszabb időszakban időbeni sorrendben állnak rendelkezésre. Az lm függvénnyel nézzük is meg a regressziót:

```
Call:
lm(formula = X[, 2] ~ k)

Coefficients:
(Intercept)            k
  1.850e-03    9.847e-09
```

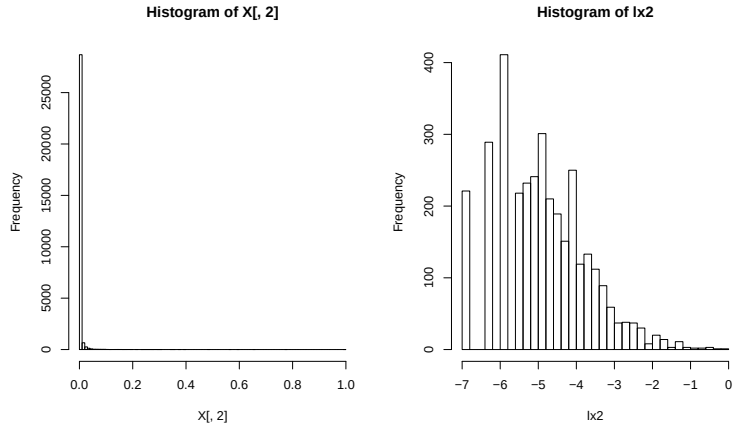
A lineáris függvény meredeksége megközelítőleg 10^{-8}, amiből arra lehet következtetni, hogy időbeni trend nincs, ha mégis akkor az a hatás elhanyagolható. Nézzük meg az ábrát is:



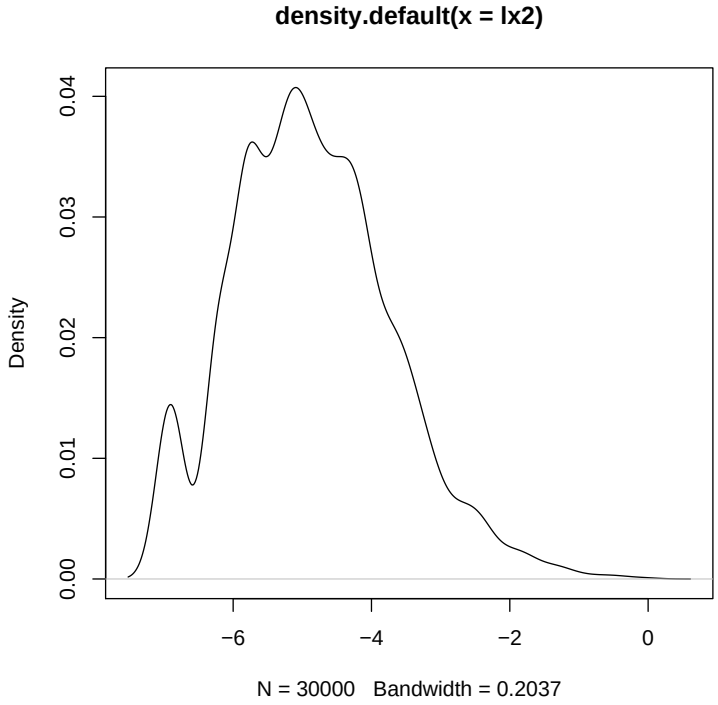
A vízszintes tengelyen a meredekség alakulása látható. Ez gyakorlatilag konstans függvényt ad eredményül, tehát levonaható a következtetés, hogy az adatok egy adott időben keletkeztek.
Nézzük még meg, hogy egy adott kárszámkategóriában mennyi az átlagos kárnagyság. Itt arra vagyunk kíváncsiak, hogy a kárszám melletti kárnagyság az összes kárra vonatkozik, vagy csak egyre.



Az első ábrán látszik egy pozitiv trend. A második ábrán, az adott kárszámmal osztott értékek találhatóak, amelyre egy konstans egyenes illeszthető. Tehát az adott kárszámokhoz összesített kárnagyságok tartoznak. Utolsóként nézzük meg, hogy alakul a kárnagyság hisztogram segítségével.



Az első ábrán látszik, hogy túl nagy a lecsengés, ha maradunk az eredeti skálán. Ezért vettem a logaritmus skálát és így néztem meg a kárnagyságokat.
Mivel jobban szeretnénk egy görbét látni, ezért a diszkrét értékekeket Parzen-Rosenblatt módszer segítségével tegyük folytonossá, a magfüggvényes sűrűségfüggvénybecslés segítségével.
Nézzünk egy illesztést a density parancs segítségével:



Log-normális eloszlás sűrűségfüggvénye rajzolódik ki.

3.1.1. Kiugró értékek kezelése

A kárnagyságnál fontos szempont a kiugró értékek meghatározása. Egy-egy nagy kár nagy mértékben eltorzíthatja a következtetéseket. Az előbb leírtakból kitűnt, hogy a 0 kárnagyság a minta több mint 80 %-ban jelen van. Ez a nagy arány a kiugró értékek kiszűrésére tett kezdeményezést meghiúsítja, a mutatók a 0 kárnagyság köré koncentrálódnak. Tehát ahhoz, hogy a szemmel látható nagy értékeket ki tudjam zárni, ahhoz először a 0 kárnagyságot veszem el. A redukált vektort rx2-nek nevezem el.

```
> rx2 <- rm.outlier(X[,2],opposite=TRUE)
```

Ekkor megnézem χ^2 out valamint a Grubbs teszttel, hogy a legnagyobb érték mekkora szignifikanciaszinten mondható kiugrónak.

```
> chisq.out.test(rx2)

        chi-squared test for outlier

data:  rx2
X-squared = 535.3803, p-value < 2.2e-16
alternative hypothesis: highest value 1 is an outlier
```

```
> grubbs.test(rx2)

        Grubbs test for one outlier

data:  rx2
G = 23.1383, U = 0.8440, p-value < 2.2e-16
alternative hypothesis: highest value 1 is an outlier
```

Az 1 kárnagyság magas szignifikanciaszinten ($< 2.2^{-16}$) kiugró értéknek tekinthető. A fenti ábrából is látszik, hogy a nagy károkból kevés van. Ezért célszerű táblázatba foglalni, hogy a 100, 110 és 120 legnagyobb érték eltávolítása után a legnagyobb érték mekkora szignifikanciaszinten mondható kiugrónak.

eltávolítottak	max. érték	χ^2 out teszt	Grubbs teszt
100	0.065	1.84^{-6}	0.002896
110	0.055	1.532^{-5}	0.02418
120	0.044	0.0002117	0.3332

A további elemzés során maradhatnánk a 110 legnagyobb kár elhagyása mellett.

Azonban a kiugró értékek eltávolítása összességében félrevezető, hiszen azt elsősorban normális eloszlású káreloszlásra lehet alkalmazni. Másrészt pedig tudjuk, hogy az adatok hitelesek - tehát nem mérési hibán alapszanak. A legnagyobb károk képezik a kiadások legnagyobb részét.

3.2. Diszkrétnek vélt magyarázó változók elemzése

A 25 magyarázó változóból 14 numerikus, vélhetőleg számok kódolva, amiből az következik, hogy területi vagy minőségi ismérv van mögötte. Az alábbi táblázat mutatja, hogy az adott változóhoz, mennyi ismérvérték tartozik.

6	7	10	12	13	16	17	18	20	21	24	25	26	27
3	4	20	10	10	2	3	2	6	4	4	7	7	6

3.2.1. Függetlenségvizsgálat

Diszkrét változóknál az elméleti részben bemutatott χ^2 próbát fogom használni. Érdemes csoportosítani a torz eredmények elkerülése végett a kárszámokat, ugyanis a nagy kárszámokhoz csekély gyakoriság tartozik, ami a próba eredményességét befolyásolja. Ezért a fent bemutatott csoportosítást

fogom használni, Tehát:

$$x1 = \begin{cases} 3 & \text{ha } x1 > 3 \\ x1 & \text{egyébként} \end{cases}$$

$$x2 = \begin{cases} 0 & 0 \leq x2 < 0.1 \\ 0.1 & 0.1 \leq x2 < 0.2 \\ 0.2 & 0.2 \leq x2 < 0.3 \\ 0.3 & 0.3 \leq x2 \end{cases}$$

Teljesebb elemzés érdekében töltsük be a vcd csomagot, és így már fogjuk tudni használni a assocstats utasítást, mely a korrekciós számításokat is tartalmazza. Előnye, hogy minden fontos adatot kiír ami a függetlenség elemzésekor érdekes lehet.

Kontigenciatáblák létrehozása után nézzük meg, diszkrét változókra a függetlenség tesztet, egy-egy változóval bemutatva, a különböző parancsokkal létrehozott eredményeket. Ahol $x102$ $xd1$ és 10-es változókból álló kontigenciatábla, továbbá $x212$ $xd1$ és 21-es változókból álló tábla, $x71$ pedig $x1$ és 7-es változókból áll. Tehát fontos, hogy a 4 illetve a 2 csoportra összevont kárszámokkal végzem az elemzést, attól függően, hogy melyik ad használható eredményt. Először a chi-négyzet tesztet nézzük meg:

```
> chisq.test(x102)

        Pearson's Chi-squared test

data:  x102
X-squared = 191.8126, df = 19, p-value < 2.2e-16
```

Ez a teszt a három legfontosabb adatot tartalmazza. A Chisq a minta alapján számított a tesztstatisztika értéke. A df (angolul: degree of freedom) a szabadságfokot mutatja. Gyakorlatilag a Chisq érték csak így értlemezhető, hiszen a szabadságfok a kontigenciatáblák nagyságát mutatja. A p-value vagyis p-érték az emprikus szignifikanciaszintet mutatja, amit az elméleti részben le is írtam. Ugyancsak az elméleti részben található az asszociációs mérőszámok bemutatása. Nézzünk olyan parancsot ami ezt mérőszámot is tartalmazza: ez az assocstats.

```
> assocstats(x212)

                    X^2 df P(> X^2)
Likelihood Ratio 2422.8  3        0
Pearson          2229.2  3        0

Phi-Coefficient   : 0.273
Contingency Coeff.: 0.263
Cramer's V        : 0.273
```

Itt a Pearson féle empirikus szignfikanicaszint olyan alacsony, hogy a program nem tudja kiszámolni. Ilyen esetekben van nagy jelentősége az Phi és Cramer együtthatóknak. (Phi-Coefficient, Cramer"s V) Elemzésünket tovább lehet finomítani a summary összesítő paranccsal, ami a Chisq.test és az egyedüli assocstats parancs eredményét kapcsolja össze. Nézzünk erre két példát.

```
> summary(assocstats(x71))

Number of cases in table: 30000
Number of factors: 2
Test for independence of all factors:
        Chisq = 56.5, df = 9, p-value = 6.294e-09
                     X^2 df   P(> X^2)
Likelihood Ratio 55.858  9 8.3556e-09
Pearson          56.502  9 6.2937e-09

Phi-Coefficient   : 0.043
Contingency Coeff.: 0.043
Cramer's V        : 0.025

> summary(assocstats(x102))

Number of cases in table: 30000
Number of factors: 2
Test for independence of all factors:
        Chisq = 191.81, df = 19, p-value = 1.437e-30
                     X^2 df P(> X^2)
Likelihood Ratio 188.32 19        0
Pearson          191.81 19        0

Phi-Coefficient   : 0.08
Contingency Coeff.: 0.08
Cramer's V        : 0.08
```

Az elsőnél látható, hogy a p-érték (p-value) megegyezik az alatta lévő táblázat Pearson - P cellában lévő értékkel. A másodiknál szembetünő, hogy a kétszer is szereplő p-érték számolásánál különböző kerekítésként jelenik meg. A többi változóra is elvégezve az elemzéseket, majd azt az alábbi táblázatba foglalva láthatjuk az eredményeket.

változó	szabadságfok	χ^2	p-érték	Cramer együttható
6	2	31.999	1.1257^{-07}	0.033
7	9	56.502	6.2937^{-09}	0.025
10	19	191.81	1.437^{-30}	0.08
12	9	77.231	5.7243^{-13}	0.051
13	9	153.83	1.418^{-28}	0.072
16	3	11.080	0.0113	0.019
17	2	238.17	1.917^{-52}	0.089
18	3	3.3626	0.339	0.011
20	5	19.154	0.001799	0.025
21	3	2229.2	0	0.273
24	9	219.64	2.489^{-42}	0.049
25	6	7.987	0.239	0.016
26	18	94.81	1.96^{-12}	0.032
27	5	77.2	3.226^{-15}	0.051

Két magyarázóváltozónál nem tudjuk elutasítani a függetlenséget. A 18-as változónál 0.339 a 25-ösnél 0.239 a p érték.

Nézzük meg azokat a változókat amelyeknél a p érték a legkisebb, hiszen a későbbi elemzés során ezek a változók lesznek az elemzés szempontjából érdekesek. Ugyanis itt magas szignifikanciaszinten el tudjuk utasítani a függetlenséget. Ezek a változók 10, 13, 17, 21, 24.

21-es változónál a p-értékre az R 0-t ír, ebben az esetben a becsült függőségre a Phi együttható ad iránymutatást. Látjuk, hogy a többihez képest magas 0.273-es értéket ad, szignifikánsan magasabb a többinél, tehát a 21. változó mutatja a legerősebb összefüggést a kárszámmal.

3.2.2. Logisztikus regresszió

Az elméleti részben részletesen bemutattam a logisztikus regressziót, mely a diszkrét adatok, azon belül is elsősorban a dichotom változók becslésére szolgál. Az R-ben a glm paranccsal lehet ezt megtenni, ahol a családot binomiálisra kell állítani. Nézzünk is egy példát!

```
Call:  glm(formula = xd1 ~ X[, 10], family = binomial())

Coefficients:
(Intercept)      X[, 10]
   -2.09607      0.01245

Degrees of Freedom: 29999 Total (i.e. Null);  29998 Residual
Null Deviance:            22060
Residual Deviance: 22040        AIC: 22050
```

Az α értéke -2.09607 mely a konstanst jelenti. A β megmutatja, hogy a jelenlegi magyarázó változó egységnyi növelésével a log odds mennyivel változik, értéke 0.01245. Ez kezdetnek nem rossz, azonban egy részletesebb elemzés, ami a hibahatárokat, szignifikanciaszintet mutatja jobb lenne. Megoldás a summary parancs.

```
> summary(glm(xd1 ~ X[,10], family=binomial()))

Call:
glm(formula = xd1 ~ X[, 10], family = binomial())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5412  -0.5196  -0.4988  -0.4844   2.0981

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.096069   0.031123 -67.348  < 2e-16 ***
X[, 10]      0.012446   0.002912   4.274 1.92e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22059  on 29999  degrees of freedom
Residual deviance: 22041  on 29998  degrees of freedom
AIC: 22045

Number of Fisher Scoring iterations: 4
```

Az első részben a reziduálisok eloszlása figyelhető meg. (Deviance Residuals) A második részben az együtthatók becslése és a hozzájuk tartozó hipotézisvizsgálatok eredménye található (Coefficients). Érdekessége, hogy nem csak a p-érték, hanem egy jelölés is van, ami a szignifikancia szintre ad egy gyors áttekinést. Ez több változó esetén lehet hasznos. Az előbbi fejezetben arra a következtetésre jutottam, hogy 5 változót érdemesebb részletesen is elemezni, a magas szignifikanciaszint miatt. Tegyük is meg az eredményt foglaljuk táblázatba.

változó	β_0	β_1	β_0 p-értéke	β_1 p-értéke
10	-2.096069	0.012446	$< 2^{-16}$	1.92^{-05}
13	-1.764631	-0.045998	$< 2^{-16}$	5.22^{-14}
17	-0.64397	-1.25351	7.85^{-12}	$< 2^{-16}$
21	0.58180	-1.90354	$< 2^{-16}$	$< 2^{-16}$
24	-1.57913	-0.29483	$< 2^{-16}$	$< 2^{-16}$

Így már az együtthatókra vonatkozó megbízhatósági szintre is választ kap-

tunk. Mindegyiknél megállapítható a magas szignifikanciaszint. Azonban ezeknek az együtthatóknak az értelmezése nem szerencsés, ezért alakítsuk át őket.

```
> exp(coef(glm(xd1 ~ X[,10], family=binomial(logit))))

(Intercept)     X[, 10]
  0.1229388   1.0125241
```

Tahát a β megmutatja, hogy a jelenlegi magyarázó változó egységnyi növelésével az odds mennyivel változik. Mondhatjuk azt is, hogy a kár bekövetkezésének esélye mennyivel nő meg. Mivel, hogy mintáról van szó, ezért a pontbecslés helyett érdemes egy konfidenciaintervallumot is meghatározni.

```
> exp(confint.default(glm(xd1 ~ X[,10], family=binomial(logit))))

                2.5 %    97.5 %
(Intercept) 0.1156637 0.1306714
X[, 10]     1.0067615 1.0183196
```

Az együtthatók tehát 95%-os konfidenciaintervallumon ilyen határok között mozognak.

A táblázatban szereplő változók szignifikanciaszintje pontosításra szorul, ugyanis a kvantilisek tanulmányozása arra enged következtetni, hogy vannak kiugró értékek. Bontsuk szét további úgynevezett dummy változókra. Ezzel a bontással egy-egy ismérvérték által gyakorolt hatást ki lehet mutatni. Ezt az elméleti részben bemutatott dummy változók bevezetésével teszem meg. Nézzük meg, hogy ennek tükrében hogyan alakul a 13-as változó:

```
> summary(glm(xd1 ~ xf13))

Call:
glm(formula = xd1 ~ xf13)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.1440  -0.1406  -0.1312  -0.1012   0.9312

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.121951   0.016898   7.217 5.45e-13 ***
xf132        0.018646   0.017240   1.082  0.27945
xf133        0.009217   0.017388   0.530  0.59605
xf134        0.019601   0.027689   0.708  0.47901
xf135       -0.034892   0.023097  -1.511  0.13088
```

```
xf136       -0.015224   0.018807  -0.809  0.41824
xf137       -0.020778   0.017789  -1.168  0.24281
xf138       -0.053136   0.017914  -2.966  0.00302 **
xf139       -0.015187   0.017725  -0.857  0.39157
xf1310       0.022021   0.018357   1.200  0.23030
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1053707)

    Null deviance: 3176.4  on 29999  degrees of freedom
Residual deviance: 3160.1  on 29990  degrees of freedom
AIC: 17640

Number of Fisher Scoring iterations: 2
```

Valóban egy-egy ismérvérték torzítja jelentősen a szignifikanciaszintet.

Illesztések után érdemes tesztelni, az 5 kiválaszott magyarázó változó együtthatóinak együttes megbizhatóságát. Likelihood hányados próbával ellenőrizzük.

$$H_0 = \beta_1 = \cdots = \beta_5$$

```
> legerosebb5 <- glm(xd1 ~ X[,10] + X[,13] + X[,17] + X[,21]
+ +X[,24], family=binomial(logit))
> red.legerosebb5 <- glm(xd1 ~ 1, family=binomial)
> anova(red.legerosebb5,legerosebb5,test="Chisq")

Analysis of Deviance Table

Model 1: xd1 ~ 1
Model 2: xd1 ~ X[, 10] + X[, 13] + X[, 17] + X[, 21] + X[, 24]
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     29999      22059
2     29994      19358  5   2701.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Magas szignifikanicaszinten elutasítható a nullhipotézis. Megjegyzem, hogy a dummy változók bevezetésekor is hasonló eredményt kapunk.

Összehasonlításként nézzük meg, hogy az öt változó együttesen milyen hatással vannak a kár-valószínűségre:

```
> summary(legerosebb5)
```

```
Call:
glm(formula = xd1 ~ X[, 10] + X[, 13] + X[, 17] + X[, 21] + X[,
    24], family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8098  -0.6883  -0.3134  -0.2190   3.8399

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.545352   0.115256  13.408  < 2e-16 ***
X[, 10]      0.013327   0.003149   4.231 2.32e-05 ***
X[, 13]     -0.019211   0.006531  -2.941  0.00327 **
X[, 17]     -0.565043   0.093096  -6.069 1.28e-09 ***
X[, 21]     -1.857640   0.047638 -38.995  < 2e-16 ***
X[, 24]     -0.316591   0.022917 -13.815  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22059  on 29999  degrees of freedom
Residual deviance: 19358  on 29994  degrees of freedom
AIC: 19370

Number of Fisher Scoring iterations: 6
```

Érdekes, hogy némiképp más az eredmény az egyenkénti elemzéshez képest a 10-es és a 13-as változónál a p érték sorrendje felcserélődött. Ez a változók egymásra gyakorolt hatásai miatt történt. Ennek megállapítására nézzük meg dummy változókkal kibővítve:

```
> summary(glm(xd1 ~ xf10 + xf13 + xf17 + xf21 + xf24,
+ family=binomial(logit)))

Call:
glm(formula = xd1 ~ xf10 + xf13 + xf17 + xf21 + xf24, family = binomial(logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9720  -0.6465  -0.3030  -0.2112   3.1818

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.55412    0.17355  -8.955  < 2e-16 ***
```

```
xf102        0.14271    0.10488   1.361 0.173603
xf103        0.31307    0.11147   2.809 0.004976 **
xf104        0.64178    0.10377   6.185 6.22e-10 ***
xf105        0.59355    0.08134   7.297 2.93e-13 ***
xf106        0.37046    0.08803   4.209 2.57e-05 ***
xf107        0.17797    0.11084   1.606 0.108347
xf108       -0.29109    0.14416  -2.019 0.043464 *
xf109        0.27509    0.09639   2.854 0.004316 **
xf1010       0.42440    0.12480   3.401 0.000672 ***
xf1011       0.24884    0.12378   2.010 0.044391 *
xf1012       0.50087    0.13050   3.838 0.000124 ***
xf1013       0.03872    0.07703   0.503 0.615185
xf1014       0.15382    0.12723   1.209 0.226670
xf1015       0.26142    0.10126   2.582 0.009837 **
xf1016       0.14486    0.19018   0.762 0.446243
xf1017       0.62892    0.12560   5.007 5.52e-07 ***
xf1018       0.40697    0.11178   3.641 0.000272 ***
xf1019       0.22862    0.13682   1.671 0.094720 .
xf1020       0.74208    0.10207   7.270 3.60e-13 ***
xf132        0.20522    0.16949   1.211 0.225959
xf133        0.23227    0.17118   1.357 0.174814
xf134        0.30755    0.26378   1.166 0.243651
xf135       -0.35835    0.24355  -1.471 0.141189
xf136        0.20073    0.19196   1.046 0.295694
xf137        0.09482    0.17738   0.535 0.592942
xf138       -0.26712    0.18372  -1.454 0.145959
xf139        0.15911    0.17637   0.902 0.366990
xf1310       0.12316    0.18058   0.682 0.495233
xf172       -0.45721    0.10409  -4.392 1.12e-05 ***
xf173       -0.16063    0.39954  -0.402 0.687655
xf212       -1.97081    0.05043 -39.078  < 2e-16 ***
xf213       -2.51219    0.22798 -11.019  < 2e-16 ***
xf214       -2.92086    0.58170  -5.021 5.13e-07 ***
xf242       -1.04766    0.16962  -6.177 6.55e-10 ***
xf243       -0.71762    0.07500  -9.569  < 2e-16 ***
xf244       -0.80636    0.08407  -9.592  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22059  on 29999  degrees of freedom
Residual deviance: 19091  on 29963  degrees of freedom
AIC: 19165
```

```
Number of Fisher Scoring iterations: 6
```

Itt is megfigyelhető, hogyha csak egy adott változót nézünk akkor teljesen eltérő eredményeket is kaphatunk, vagy ha többet nézünk, de nem vezetjük be a dummy változókat akkor is ilyen eltérő eredményeket kapunk. Például a 13 változó egyik ismérvértéke sem szignifikáns, ezért a díjkalkulációval foglalkozó 4. fejezetben a legszorosabb kapcsolatot mutattó dummy változókat vesszük figyelembe.

3.3. Folytonosnak vélt változók elemzése

E fejezetben ugyancsak a kapcsolat szorossága, valamint a kiugró értékek feltárása lesz a fő irányvonal.

3.3.1. Korrelációszámítás

Az ismérvértékekből tudok következtetni, hogy a változók nem nominális skálán mozognak, tehát lehet ordinális intervallum vagy arányskála. Mivel, hogy egyéb rendelkezére álló adatom nincs ezért abból indulok ki, hogy ordinális skálán mozognak a változók értékei. Az előbb leírtak ismeretében a rangkorreláció elemzésével kezdem. A kapcsolat szorosságát a rendelkezésre álló mintából tesztelem. Az R-ben erre a cor.test parancs a legalkalmasabb, mellyel nemcsak a Spearman és Kendall, hanem a Pearson korreláció is elvégezhető. Utaltam arra, hogy mivel nem tudom pontosan milyen skálán mozgok ezért rangkorrelációt alkalmazok. Ennek ellenére megnézem összehasonlítás gyanánt a Pearson-féle korrelációval is az összefüggés szorosságát. Nézzük meg a 22-es magyarázóváltozóra.

```
        Pearson's product-moment correlation

data:  X[, 2] and X[, 22]
t = 18.017, df = 29998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.09225847 0.11464810
sample estimates:
      cor
0.1034664

        Spearman's rank correlation rho

data:  X[, 2] and X[, 22]
S = 3.572371e+12, p-value < 2.2e-16
```

```
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.2061397

        Kendall's rank correlation tau

data:  X[, 2] and X[, 22]
z = 35.8987, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.1647203

[1] 3636
```

A statisztikákon túlmenően engem igazából két érték érdekel egy-egy tesztből. Az egyik a korábban már megismert p-érték, mely itt is a szignifikanciaszintet határozza meg. Vagyis a $\rho = 0$ hipotézis elutasításának megbízhatóságára ad választ. A másik pedig a szorosságot kifejező érték.

A többi változóról kapott eredményt egy táblázatba foglalom, a könnyebb áttekintés miatt. Spearman teszt p-értékét nem írom bele, mert gyakorlatilag megegyezik a Kendall teszthez tartozó p-értékkel.

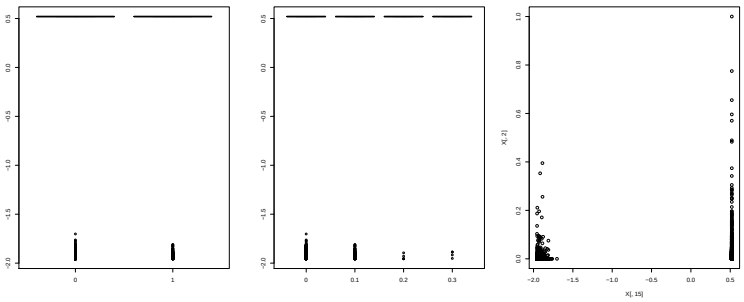
változó	pearson	p értéke	spearman	kendall	p értéke
3	-0.015532	0.007139	-0.021251	-0.016986	0.000213
4	-0.014972	0.009509	-0.012019	-0.009650	0.0375
5	0.060427	$< 2.2^{-16}$	0.151459	0.120515	$< 2.2^{-16}$
8	-0.018072	0.001745	0.021977	0.017227	0.000177
11	-0.002956	0.6086	0.028481	0.023285	7.778^{-7}
14	-0.044265	1.715^{-14}	-0.054166	-0.043622	$< 2.2^{-16}$
15	0.027605	1.734^{-06}	0.078136	0.072178	$< 2.2^{-16}$
19	0.012701	0.02781	0.047983	0.041751	$< 2.2^{-16}$
22	0.103466	$< 2.2^{-16}$	0.206140	0.164720	$< 2.2^{-16}$
23	0.121223	$< 2.2^{-16}$	0.244814	0.195477	$< 2.2^{-16}$

Több észrevételem is van a táblázatban szereplő értékekkel kapcsolatosan. Az egyik az, ami talán a legfontosabb, hogy még a legszorosabb kapcsolatot mutató 23-as változó is csak 0.195477 Kendall-féle és 0.244814 Spearman-féle szorossági együtthatóval rendelkezik, persze magas szignifikanciaszinten. Kendall rangkorrelációnál a p-értékek megfelelőek. Megfigyelhető, hogy a Pearson-féle korreláció együtthatói kissebbek mint a Kendall-féle rangkorreláció együtthatói. Ez abból adódhat, hogy a rangkorreláció kevésbé érzékeny a szélsőséges értékekre. Így adódik a következő rész témája a kívülálló értékek kezelése.

3.3.2. Kiugró értékek megállapítása

Az értékek megállapítása több tényezőből tevődik össze, nincs egzakt módszer. Egy egyszerű table illetve plot parancs segítségével az eloszlásról kaphatunk képet. Célszerű a kárnagyság kategóriái szerinti eloszlást is megnézni, boxplot függvény segítségével. Függvénynél látszik az alsó, felső kvartilis medián valamint a legkisebb és legnagyobb érték is. Azonban ezek a kvantilisek összemosódnak. Ezen kívül marad az elméleti részben is bemutatott tesztek futattása. Megnézem, hogy a 15-ös változónál ez hogyan is néz ki.

```
> length(table(X[,15]))

[1] 111

> chisq.out.test(X[,15])

        chi-squared test for outlier

data:  X[, 15]
X-squared = 3.8259, p-value = 0.05046
alternative hypothesis: lowest value -1.956 is an outlier

> grubbs.test(X[,15],type=10)

        Grubbs test for one outlier

data:  X[, 15]
G = 1.9560, U = 0.9999, p-value = 1
alternative hypothesis: lowest value -1.956 is an outlier

> grubbs.test(X[,15],type=10, opposite= TRUE)

        Grubbs test for one outlier

data:  X[, 15]
G = 0.5201, U = 1.0000, p-value = 1
alternative hypothesis: highest value 0.52 is an outlier

> grubbs.test(X[,15],type=11)

        Grubbs test for two opposite outliers

data:  X[, 15]
G = 2.4761, U = 0.9999, p-value = 1
alternative hypothesis: -1.956 and 0.52 are outliers
```



A length utasítás az ismérvértékek számát jelenti ennél a változónál 111. A χ^2 és Grubbs teszt, a statisztikákon kívül a legkisebb, (-1.956) illetve a legnagyobb (0.52) kiugró értékhez tartozó empirikus p értéket is meghatározza. Összeségében a tesztek eredményei azt sugallják, hogy nincs kiugró elem, azonban a tábla és a grafikák tanulmányozása azt mutatja, hogy van 0.52 ilyen érték.

Táblázatba foglalom a többi változóhoz tartozó értékeket.

változó	értékek száma	legkisebb érték	legnagyobb érték	χ^2 out felső	Grubbs felső	Grubbs alsó	Grubbs szélek
3	3450	-1.715	1.735	0.08274	1	1	1
4	88	-2.585	6.769	1.3^{-11}	1.916^{-7}	1	0.01657
5	1094	-1.568	1.94	0.05238	1	1	1
8	1474	-1.258	2.035	0.04185	1	1	1
11	66	-1.958	0.561	0.05017	1	1	1
14	325	-6.502	1.802	7.929^{-11}	1.172^{-6}	1	1
15	111	-1.956	0.52	0.05046	1	1	1
19	11	-1.705	5.183	2.19^{-7}	0.003265	1	1
22	3636	-0.977	31.91	$< 2.2^{-16}$	$< 2.2^{-16}$	1	$< 2.2^{-16}$
23	3485	-0.905	36.594	$< 2.2^{-16}$	$< 2.2^{-16}$	1	$< 2.2^{-16}$

Mindent egybevetve a 4-es 11-es 14-es 15-ös 19-es 22-es 23-as változóknál szükséges a kiugró értékek kezelése, eltávolítása. Eljárásom célja, hogy a redukált adathalmazra korrelációs teszteket tudjak felírni. Ez a 11 és 15 változón nem segít, mert egy ismérvérték eltávolítása az elemszámot jelentősen csökkenti. Ebben az esetben az adott magyarázóváltozó és eredményváltozó között korrerálatlanság keletkezik. A teszt alapján kiugró értékek eltávolítása után sem javult a korrelációra irányuló tesztem. Ilyen szorossági értékek mellett a regressziószámítás sem ad megbizható eredményt. Így a kiugró értékek elemzését a továbbiakban mellőzöm, az előző hasonló rész konzekvenciája miatt.

4. fejezet

Díjkalkuláció

A biztosítótársaság számára fontos a megfelelő díj meghatározása, hogy biztonságosan tudjon működni. A díjak meghatározása az elméleti részben bemutatott díjkalkulációs elvek segítségével történik. Előtte a minta alapján becsüljük meg a sokaság várható értékét szórását.

4.1. Káreloszlás jellemző értékei és azok becslése

Először nézzük meg a kárnagyság minta alapján becsült jellemzőit:

tapasztalati várható érték	$tapvarert = 0,0019979$
tapasztalati szórásnégyzet	$tapszornegyzet = 0.0002373$
tapasztalati szórás	$tapszor = 0.0154038$

Az adatelemzési részben megállapítottam, hogy a káreloszlás a lognormális eloszláshoz közelít. Az eloszlás paramétereit momentum módszerrel is becsülhetjük, amit a díjszámításnál is fel tudok használni.

Másrészt van egy 30000-es mintánk melynek várható értéke és szórása mintáról mintára változik.

Általánosságban megállapítható, hogy ezek az értékek normális eloszlást alkotnak. [1] 90%-os megbízhatóságú konfidencia intervallum felső végpontja az összkárra:

```
> elmosszkar <- osszkar + qnorm(0.95)*sqrt(30000)*tapszor
```

Az eredmény az elméleti összkárra illetve az egyedi kárnagyságra:

```
[1] 64.3255
```

```
[1] 0.3713835
```

[1] Marits Ágnes: A kockázati díjak kalkulációja a kárbiztosításban 42. old. alapján

4.2. Díjak meghatározása különböző díjelvek segítségével

A lognormális eloszlás várható értékét és szórását (itt jelöljük: μ ésσ) felhasználva a következő összefüggéssel kiszámolható a különböző valószínűség mellett a szolgáltatásért cserébe elkérhető díj.

$P(Z > EZ + a * DZ) = 1 - \Phi(k)$ Ahol $k = \frac{\ln a - \mu}{\sigma}$, $a =$ arányossági tényező

Az aggregált kockázatot - különböző szignifikanciaszintek mellett - a következő arányossági tényezők (pl. $lkar95$) segítségével tudom meghatározni:

```
> lkar95 <- exp(qnorm(0.95)*szoras+varert)
```

A várható érték a szórásnégyzet és a szórás elvvel az elméleti rész alapján a következő aggregált kockázatok határozhatóak meg:

	várható érték elv	szórás elv	szórásnégyzet elv
95%	64.78792	64.32561	64.3255

Megfigyelhető, hogy a legdurvább becslés a várható érték elvvel, a legjobb becslés a szórásnégyzet elvvel valósítható meg, a kockázat kiszámítására. Nézzük meg, hogy ez mit jelent egy kárra lebontva:

	várható érték elv	szórás elv	szórásnégyzet elv
95%	0.002159597	0.002144187	0.002144183

Az előző fejezetben bemutatott logisztikus regresszióval, és a dummy változók segítségével meg tudom határozni, a különböző tulajdonságú szerződők díjait. Fontos, hogy csak megfelelő szignifikanciaszintű változókkal lehet számolni, ez ad megbízható eredményt. Válasszuk ki a három csillagosokat. Másrészt dummy változókról beszélünk tehát x_i 0 vagy 1. Így az elméleti részben leírtak alapján könnyen kiszámolható az odds (pl: $e10$) és ebből a kár bekövetkezésének valószínűsége. Az 5 különböző tulajdonságok közül válasszunk ki egy-egy ismérvet. Egyszerre csak egy ismérv egy tulajdonságának a valószínűségét nézzük. A különböző értékek a különböző együtthatók (ismérvváltozatok) valószínűségét jelelnti.(Pl. $p10[5]$)

```
> e10 <- exp(coef(glm(xd1 ~ xf10, family=binomial(logit))))
> p10 <- e10/(1+e10)
> p10[5]

    xf105
0.6554938
```

Az egy szerződőre eső díj az előbbi táblázatban megtalálható, ez lesz a kiindulópont. A tulajdonsághoz tartozó valószínűséggel szorozva kapjuk a kedvező csoport egyedi díját. Így az adott csoportra számolható egy összkár. A teljes összkárból kivonva a kedvező csoport összkárát, majd osztva a maradék szerződőre, megkapom a többi tulajdonsággal rendelkezőre elkérhető díjat. Az utolsó oszlop tehát a komplementer csoporthoz tarozó egyedi díjakat mutatja. Nézzük az alábbi táblázatot.

tulajdonság	ismérv	csoport mérete	valószínűsége	egyedi díja
10	5	2218	0.6554938	0.001405499
13	1	369	0.1219512	0.0002614858
17	2	3138	0.2078885	0.000445751
21	2	14399	0.12055763	0.0002584977
24	4	2664	0.3146417	0.0006746496

Fontos az elemszám is. Pl. a 21 változó 2 ismérvértéke nagy számban előfordul és ehhez alacsony valószínűség tartozik. A díjkülönbség ennél a változónál a legszembetűnőbb. Megvalósul a nagy tömegek elérése, kedvező díjjal, amire a biztosító különböző stratégiákat építhet.

Több tulajdonság egy-egy ismérvének együttes valószínűsége az, ami talán a legjobban érdekelheti a biztosítótársaságot. Olyan tulajdonságokat kell összeválogatni, ami magas szignifikanciaszint mellett a kár alacsony valószínűséggel következik be. Nézzünk egy példát erre:

```
> evalt <- exp(coef(glm(xd1 ~ xf10 + xf13 + xf17 + xf21 + xf24,
+ family = binomial(logit))))
```

Ebből megkaptuk a megfelelő esélyhányadosokat, melyből adódik a megfelelő valószínűség.

```
> sevalt <- evalt[5]*evalt[30]*evalt[32]*evalt[37]
> pvalt <- sevalt/(1+sevalt)
> pvalt
```

```
    xf105
0.0665569
```

Ezekkel a tulajdonságokkal kellően alacsony díjakat tudok meghatározni:

```
> (elmosszkar + tapszornegyzet*lkar95)/30000*pvalt
```

```
       xf105
0.0001427102
```

Az összeválogatásnál fontos az előbb is kihangsúlyozott elemszám. A példában szereplő kombináció elemszáma 105. Így, ezen tulajdonságokkal rendelkező csoport összdíjának értéke 0.01498457 a komplementeré pedig 0.002151213.

5. fejezet

Összefoglalás

A biztosítótársaság érdeke alacsony kockázatú szerződések megkötése. Természetesen akkor van előnyben, ha nagy biztonsággal állíthatja, hogy a rendelkezésre álló adatok alapján ez így is fog történni. A dummy változók alkalmazásával, a nominális skálán mért tulajdonságok jól kezelhetővé váltak. A logisztikus regresszió pedig a kár bekövetkezésének valószínűségét határozza meg. Ezen adatok felhasználásával a különböző díjkalkulációs elvek segítségével a konkrét díjakat tudtunk meghatározni. Több tulajdonság együttes elemzésével, magas megbízhatósággal még alacsonyabb díjak határozhatóak meg.

Ugyan a dolgozatom témája biztosítói káradatok elemzése volt, azonban ezen módszerek más gazdasági területen is jól alkalmazhatóak mint. pl. csődkockázat vagy hitelbírálat. Itt is kockázatok (csőd, hitel vissza nem fizetése) valószínűségét kell becsülni, ami hasonlóan történhet, mint ebben a dolgozatban. A feladatot nagyban nehezítette a biztosító azon kérése, hogy a változók tényleges jelentését sem ismerhettük meg.

Irodalomjegyzék

[1] Arató Miklós: Nem-életbiztosítási matematika, 2001

[2] Marits Ágnes: A kockázati díjak kalkulációja a kárbiztosításban, *MKKE Biztosítási kutató csoport biztosításelméleti füzetek 4.*, 1988. december

[3] Király Zoltán: Statisztika II,
http://psycho.unideb.hu/munkatarsak/hidegkuti_istvan/targyak/Kiraly_Zoltan_Statisztika_2_jegyzet_1.pdf
http://psycho.unideb.hu/munkatarsak/hidegkuti_istvan/targyak/Kiraly_Zoltan_Statisztika_2_jegyzet_2.pdf

[4] Michael Friendly: Working with categorical data with R and the vcd and vcdExtra packages, *York University, Toronto*, 2013
http://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf

[5] Csicsman József - Sipos Szabó Eszter: Matematikai alapok az adatbányászati szoftverek első megismeréséhez,
http://www.inf.u-szeged.hu/~csicsman/oktatas/kornyezettan/Fuggelek/stat_book.pdf

[6] Ferenci Tamás: Ökonometria, Logisztikus regresszió, *Budapesti Corvinus Egyetem*
http://www.medstat.hu/oko/2011osz/eloadas8slides.pdf

[7] Orlovits Zsanett: Nominális változók a lineáris modellben *BME*
http://www.math.bme.hu/~orlovits/GPK_SZTOCH_EA_REG3.pdf

[8] Ferenci Tamás: Ökonometria, Dummy változók használata, *Budapesti Corvinus Egyetem*
http://www.medstat.hu/oko/2011osz/eloadas7slides.pdf