UNIVERZA V LJUBLJANI
FAKULTETA ZA MATEMATIKO IN FIZIKO

Finančna matematika – 2. stopnja

Doris Klančar

# Uporaba posplošenega linearnega modela v zavarovalništvu

Magistrsko delo

Mentor: prof. dr. Tomaž Košir
Somentor: dr. Janez Komelj

Ljubljana, 2015

Podpisana Doris Klančar izjavljam:

- da sem magistrsko delo z naslovom *Uporaba posplošenega linearnega modela v zavarovalništvu* izdelala samostojno pod mentorstvom prof. dr. Tomaža Koširja ter dr. Janeza Komelja in
- da Fakulteti za matematiko in fiziko Univerze v Ljubljani dovoljujem objavo elektronske oblike svojega dela na spletnih straneh.

Ljubljana, 3. 7. 2015 Podpis: .....

# Zahvala

Zahvaljujem se mentorju, prof. dr. Tomažu Koširju, za strokovno pomoč in nasvete pri izdelavi magistrskega dela.

Hkrati se zahvaljujem somentorju, dr. Janezu Komelju, za vse nasvete in vodenje pri pisanju dela. S svojimi izkušnjami in strokovnim znanjem o temi mi je bil v veliko pomoč.

Posebna zahvala gre moji družini in prijateljem, ki so me vzpodbujali v času celotnega študija.

# Kazalo

# Program dela

V magistrskem delu opišite posplošeni linearni model in njegovo uporabo v zavarovalništvu. Predstavite tudi konkreten zgled.

Mentor: prof. dr. Tomaž Košir

Ljubljana, 1. 10. 2014

Somentor: dr. Janez Komelj

# Povzetek

V magistrskem delu predstavljamo uporabo posplošenega linearnega modela – GLM (generalized linear model) v zavarovalništvu. V GLM posplošimo predpostavke linearnega modela, kjer je ključna posplošitev ta, da slučajna spremenljivka, ki jo modeliramo, ni nujno normalno porazdeljena, ampak pripada družini eksponentno porazdeljenih slučajnih spremenljivk. GLM je zaradi tega zelo uporaben pri oblikovanju cen posameznih zavarovanj, saj v to družino spadata tudi Poissonova in gama porazdelitev, ki ju v zavarovalništvu pogosto uporabljamo za modeliranje škodne pogostosti in povprečne škode. Hkrati v GLM posplošimo tudi predpostavko, da naj bi bila modelirana slučajna spremenljivka linearno odvisna od neodvisnih spremenljivk, kar nam omogoča uporabo multiplikativnega modela, ki je v zavarovalništvu zelo uporaben.

Ker gre za posplošen linearni model, v delu najprej predstavljamo osnovne predpostavke linearnega modela. Nato je predstavljen GLM, kjer je prikazano, kako ocenimo parametre in tudi kako lahko postavljeni model ovrednotimo. S testiranjem hipotez namreč ocenimo, kako dobro se postavljeni model prilega podatkom, hkrati pa lahko tudi primerjamo modele med seboj. Omenjene so tudi možne nadgradnje in dopolnitve osnovnega modela. Na koncu je prikazan tudi preprost primer, kjer smo modelirali škodno pogostost in povprečno škodo, vsi izračuni pa so narejeni v programskem orodju R.



# Abstract

In the thesis we present the use of generalized linear model (GLM) in insurance. In the GLM we generalize the assumptions of the linear model. The key assumption of the generalized linear model is that the random variable is not necessarily normally distributed, but it belongs to the family of exponentially distributed random variables. Therefore, the GLM is very useful for modelling of prices of individual insurances, since the Poisson and gamma distribution are also a part of this family, and are often used for modelling claim frequency and claim severity in insurance. At the same time in the GLM, we also generalize the assumption which states that the formed random variable should be linearly dependent on independent variables. That enables us the use of multiplicative model, which is very useful in insurance.

Since we present the generalized linear model, first we introduce the basic assumptions of the linear model. Further, the GLM is introduced, where the way of estimating the parameters as well as the way of evaluating of the model is presented. By testing the hypotheses, we evaluate how the introduced model fits the data, and at the same time we can also compare the individual models. There are also the possible upgrades and completions of the basic model mentioned. In conclusion, there is also a simple example presented, in which we are modelling claim frequency and claim severity. All the calculations are made in software tool R.



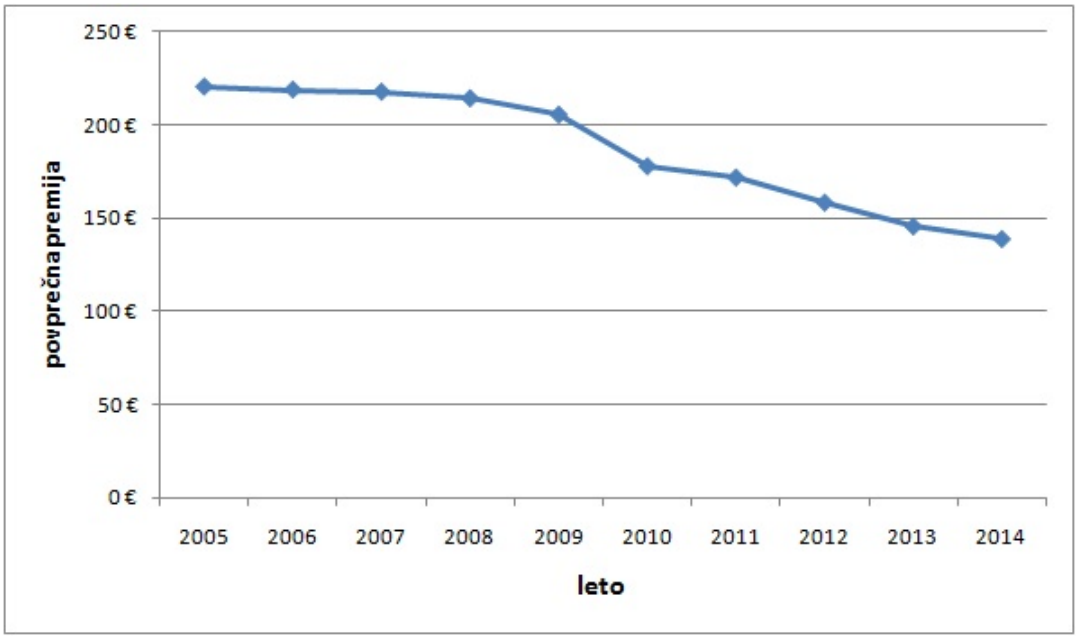
**Math. Subj. Class. (2010):** 62J12, 62P05
**Ključne besede:** posplošen linearni model, oblikovanje cen v zavarovalništvu
**Keywords:** generalized linear model, insurance pricing

# 1 Uvod

Zavarovalništvo je v Sloveniji od začetka devetdesetih let prejšnjega stoletja pa do leta 2008 dosegalo visoke stopnje rasti, ki so bile višje od splošne rasti v gospodarstvu. Z nastopom finančne krize so zavarovalnice izkazovale slabše poslovne rezultate, zahtevne razmere pa so se nadaljevale v leto 2013, kar je negativno vplivalo tudi na slovensko zavarovalništvo, saj je upadla zbrana premija ([10, str. 9]).

Dober pokazatelj zaostrovanja razmer na zavarovalniškem trgu je statistika obveznega zavarovanja avtomobilske odgovornosti, saj predstavlja velik delež vseh zavarovanj v Sloveniji. Spodnji graf prikazuje upad povprečne premije za zavarovanje avtomobilske odgovornosti.

Slika 1: Povprečna premija zavarovanj avtomobilske odgovornosti za obdobje 2005-2014 ([6, str. 27])




Ker je na drugi strani nastala škoda skozi leta padala manj kot povprečna premija, to pomeni, da so zavarovalnice zaradi tega v težjem položaju.

Na zavarovalniškem trgu se je tako pojavila potreba po razvijanju boljših metod za oblikovanje cen posameznih zavarovanj. Če za posamezno zavarovanje velja, da povprečna premija pada, to pomeni, da moramo na drugi strani poskrbeti, da bodo na enak način padale tudi izplačane škode. To pa lahko dosežemo tako, da zavarovalnica v zavarovanje sprejme le rizike, ki so v povprečju manj rizični. V splošnem bi to pomenilo, da mora zavarovalnica razviti metodo, s katero bo na trgu prepoznala manj rizične segmente zavarovancev, ki jim bo lahko zaračunala nižjo premijo kot konkurenca, bolj rizičnim pa se bo izognila (zaračunala višjo premijo kot konkurenca). Temu rečemo tudi pobiranje smetane, kar pomeni, da želi zavarovalnica na trgu pobrati le manj rizične zavarovance.

Ključno je torej prepoznavanje tistih zavarovancev, ki so na trgu manj rizični in jim na ta račun ponuditi premijo, ki bo nižja kot tista, ki jo bomo zaračunali bolj

rizičnim zavarovancem. To dejansko pomeni, da zavarovancem zaračunamo tako premijo, kot si jo zaslužijo glede na njihove lastnosti. Eden od načinov, kako določimo tako premijo, je tudi s pomočjo posplošenega linearnega modela. S pomočjo tega modela lahko prepoznamo lastnosti, ki vplivajo na rizičnost zavarovanca, hkrati pa tudi določimo, v kolikšni meri ta lastnost vpliva na povečanje/zmanjšanje rizičnosti, torej v kolikšni meri vpliva na povečanje/zmanjšanje zahtevane premije. V svojem delu sem se zaradi tega osredotočila na uporabo posplošenega linearnega modela v zavarovalništvu. Pri tem sem se naslonila predvsem na dve deli, in sicer na delo *E. Ohlsson, B. Johansson: Non-Life Insurance Pricing with Generalized Linear Models (2010)* in na delo *D. Anderson idr.: A Practitioner's Guide to Generalized Linear Models (2007)*, ki ravno tako opisujeta uporabnost posplošenega linearnega modela v zavarovalništvu.

# 2 Uvod v zavarovalništvo

## 2.1 Zavarovanje

Največkrat se za definicijo zavarovanja uporablja opredelitev, ki jo je uporabil Boncelj v Zavarovalni ekonomiki: »Zavarovanje je gospodarska institucija ustvarjanja gospodarske varnosti z združevanjem različnih objektov zaradi izravnavanja nevarnosti.« ([4, str. 13]).

Naloga zavarovanja je, da številna tveganja, ki so jim zavarovanci izpostavljeni, prerazporedi na vse zavarovance in da zavarovancu izplača ustrezno nadomestilo za utrpelo škodo ali ustrezno vsoto v skladu s sklenjeno zavarovalno pogodbo. Za posameznika nikoli ne moremo trditi, ali bo izpostavljen uresničitvi nekega škodnega dogodka, za veliko skupino posameznikov pa pojav določenih škodnih dogodkov lahko predvidimo z veliko verjetnostjo ([3, str. 9]). Razlog, da se odločimo za zavarovanje, je v tem, da negotovost zamenjamo za gotovost, saj se z vnaprej znanim in razmeroma majhnim plačilom premije zavarujemo pred negotovim dogodkom v prihodnosti, katerega škoda je lahko zelo velika. Škodo bo namreč v primeru sklenitve zavarovanja krila zavarovalnica.

Zavarujemo se torej proti določenemu tveganju. Tveganje (riziko) v zavarovanju je možnost naključne uresničitve zavarovane nevarnosti na predmetu zavarovanja v nekem prihodnjem časovnem obdobju. Za zavarovalnico je zelo pomembna čim natančnejša določitev velikosti tveganja oziroma ovrednotenje posledic vseh nevarnosti, ki lahko nastanejo na predmetu zavarovanja. Velikost tveganja se namreč kaže v višini premije. Zavarovalna tveganja skupaj običajno imenujemo portfelj zavarovalnice. Velikost portfelja merimo z vplačano premijo, z vrednostjo ali številom zavarovanih predmetov, ali kot vsoto zavarovalnih vsot pri življenjskih zavarovanjih ([11, str. 10]).

Pomembno je, da so pri sklenitvi zavarovanja izpolnjeni tudi nekateri pogoji, ki omogočajo, da se zavarujemo samo pred negotovim dogodkom, za katerega lahko v primeru nastanka od zavarovalnice dobimo povrnjeno škodo. Določeno je, da mora biti ob sklenitvi zavarovanja zavarovalni dogodek v prihodnosti negotov in neodvisen od zavarovalčeve volje, da je tveganje mogoče ovrednotiti in je škodo mogoče oceniti. Zavarovalni dogodek ali zavarovalni primer je dogodek, ki je nastal zaradi nevarnosti, ki je po zavarovalni pogodbi pokrita in zaradi katere je nastala škoda na zavarovanem predmetu. ([11, str. 9, 86])

Pri vsakem zavarovanju je potrebno določiti:

- predmet (objekt) zavarovanja,
- tveganje, za katero je predmet zavarovanja zavarovan,
- čas trajanja zavarovanja,
- obliko škodnega kritja.

Velikost tveganja je odvisna predvsem od naslednjih dejavnikov:

- vrste tveganja (požar, poplava, avtomobilska nesreča, smrt,...),
- fizično-tehničnih značilnosti predmeta zavarovanja (vrsta avtomobila, gradnja objekta,...),
- vrednosti predmeta zavarovanja in s tem višine prevzetih obveznosti (kritja),
- trajanja zavarovanja,
- kraja, kjer se nahaja predmet zavarovanja.

([11, str. 9, 10])

## 2.2 Zavarovalna pogodba

Celotno razmerje med zavarovalnico in zavarovalcem določa zavarovalna pogodba. Določba 921. člena Obligacijskega zakonika definira zavarovalno pogodbo kot:
»Z zavarovalno pogodbo se zavarovalec zavezuje, da bo zavarovalnici plačal zavarovalno premijo ali prispevek, zavarovalnica pa se zavezuje, da bo, če se zgodi dogodek, ki pomeni zavarovalni primer, izplačala zavarovancu ali nekomu tretjemu zavarovalnino ali odškodnino ali storila kaj drugega.«

Iz definicije zavarovalne pogodbe je vidno, da se pogodba sklepa med zavarovalnico in zavarovalcem. V primeru, da se zgodi zavarovalni dogodek, pa se zavarovalnino izplača zavarovancu ali nekomu tretjemu, ki mu rečemo tudi upravičenec. Razlika med zavarovalcem, zavarovancem in upravičencem je naslednja:

- **Zavarovalec:** (sklenitelj) zavarovanja je tisti, ki sklene pogodbo in plača premijo.
- **Zavarovanec:** je oseba, katere premoženjski interes je zavarovan. Zavarovalec in zavarovanec sta pogosto ista oseba, odvisno od vsebine zavarovalne pogodbe in vrste zavarovanja. Pri osebnih zavarovanjih je to oseba, od katere smrti, invalidnost ali okvare zdravja je odvisno izplačilo zavarovalnine. Pri premoženjskih zavarovanjih odgovornosti pa je to oseba, ki bi v primeru nastanka zavarovalnega primera morala plačati odškodnino oškodovancu. Pri drugih premoženjskih zavarovanjih je to lastnik vozila (kasko zavarovanje) ali hiše (požarno zavarovanje).
- **Upravičenec:** je oseba, ki ji zavarovalnica izplača zavarovalnino oziroma odškodnino.

  ([11, str. 89, 90])

## 2.3 Premija

Zavarovalna premija je znesek, ki ga zavarovalec plača zavarovalnici ob sklenitvi zavarovanja v zameno za zavarovalno kritje. Temu znesku rečemo tudi bruto ali kosmata premija. Kosmata premija se deli na funkcionalno premijo in stroškovno

premijo, ki je namenjena za pokrivanje stroškov zavarovalnice. Funkcionalna premija se deli še naprej na tehnično premijo in dodatek za preventivo in represijo. Z dodatkom za preventivo zavarovalnica izvaja razne ukrepe, s katerimi zmanjšuje možnost nastanka škode, z dodatkom za represijo pa izvaja ukrepe za reševanje ljudi in premoženja oziroma znižuje višino škode v primeru nastanka škode. Tehnična premija pa se nato deli še naprej na nevarnostno premijo, varnostni dodatek in v primeru življenjskih zavarovanj še na varčevalno premijo. Nevarnostna premija je premija, ki v povprečju zadošča za kritje vseh škod iz tveganj, ki so bila sprejeta v zavarovanje, varnostni dodatek pa je potreben zaradi možnega negativnega odmika od povprečnega škodnega dogajanja ([7]).

Ključen del pri določanju cen v zavarovalništvu je izračun nevarnostne premije, saj mora ta del pokriti škode, ki bodo nastale iz tveganj sprejetih zavarovanj. Ob obračunu premije seveda še ne vemo, kolikšne bodo škode, zato nevarnostno premijo v praksi izračunamo na podlagi statistike preteklih škod, in sicer kot pričakovano vrednost višine škode na enoto izpostavljenosti.

Izpostavljenost je mera, s katero zavarovalnica meri, kolikšnemu tveganju je v določenem obdobju izpostavljena. Izpostavljenost se razlikuje glede na vrsto zavarovanja, pri premoženjskih zavarovanjih je to lahko kar vsota zavarovalnih vsot zavarovanj, pri nekaterih drugih zavarovanjih, kot je na primer zavarovanje avtomobilske odgovornosti, pa je izpostavljenost enaka številu enakovrstnih rizikov, ki so zavarovani v nekem časovnem intervalu, največkrat je to eno leto ([7]). Nevarnostno premijo torej izračunamo kot produkt škodne pogostosti in povprečne škode.

**Primer 2.1.** Zavarovalnica je v preteklosti sprejela v zavarovanje $n$ enoletnih tveganj (enoletne police). Izplačala je $k$ škod, ki jih označimo z $x_1, x_2, ..., x_k$. Celoten znesek, ki ga je izplačala zavarovalnica, je torej enak

$$S_n = \sum_{i=1}^{k} x_i.$$

Na podlagi teh škod lahko izračunamo:

- **škodna pogostost** – število škod na enoto izpostavljenosti

$$škodna\, pogostost = \frac{k}{n}.$$

- **povprečna škoda** – povprečna izplačana škoda

$$povprečna\, škoda = \frac{S_n}{k}.$$

- **nevarnostna premija** – produkt škodne pogostosti in povprečne škode oziroma izplačana škoda na enoto izpostavljenosti

$$\begin{aligned} nearnostna\, premija &= škodna\, pogostost \times povprečna\, škoda \\ &= \frac{k}{n}\frac{S_n}{k} = \frac{S_n}{n}. \end{aligned}$$

# 3 Oblikovanje cen v zavarovalništvu

Zavarovalnica z večanjem števila zavarovancev manjša tveganje, saj se manjša varianca povprečnega tveganja, ki ga je sprejela v zavarovanje. Na ta način zavarovalnica bolj natančno oceni pričakovano škodo, saj se odkloni od povprečja z večanjem števila zavarovanj manjšajo. Nevarnostna premija je tako enaka pričakovanim škodam, ki so posledica sprejetja tveganja v zavarovanje. To govori tudi centralni limitni izrek ali zakon velikih števil.

**Izrek 3.1.** *(Centralni limitni izrek)*
*Naj bodo $X_1, X_2, ..., X_n$ neodvisne, enako porazdeljene slučajne spremenljivke, za katere velja*

$$\mathbb{E}[X_1] = \mu,$$
$$\mathrm{VaR}[X_1] = \sigma^2.$$

*Potem velja, da se porazdelitev vsote $S_n = X_1 + ... + X_n$ približuje normalni porazdelitvi, ko gre $n \to \infty$*

$$S_n = \sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} N(n\mu, n\sigma^2),$$

$$\frac{S_n}{n} \xrightarrow{n\to\infty} N\left(\mu, \frac{\sigma^2}{n}\right).$$

Seveda pa se tveganja, ki jih zavarovalnica sprejme v zavarovanje, razlikujejo po svoji rizičnosti, kar pomeni, da bo tudi pričakovana škoda teh tveganj različna. Primer različnih tveganj so na primer mladi vozniki in vozniki z več leti vozniških izkušenj. Dejstvo je, da so mladi vozniki v povprečju zaradi neizkušenosti veliko bolj rizični kot pa bolj izkušeni vozniki. Če zavarovalnica teh razlik ne upošteva pri oblikovanju premije in vsi zavarovanci plačujejo enako premijo, to pomeni, da zavarovanci z več vozniškimi izkušnjami plačujejo višjo premijo, kot pa je njihova pričakovana škoda, saj z višjo premijo kompenzirajo mlade voznike, ki plačujejo premalo. Problem se pojavi, če bi konkurenčna zavarovalnica zaračunavala dve različni premiji za ti dve vrsti zavarovancev. Poglejmo si to na preprostem primeru.

**Primer 3.2.** Naj bo $p_1$ potrebna premija za bolj rizične zavarovance, v našem primeru mlade voznike, $p_2$ pa potrebna premija za manj rizične zavarovance – voznike z več leti vozniških izkušenj, velja torej $p_2 < p_1$. Naj bo $n_1$ število bolj rizičnih zavarovancev in $n_2$ število manj rizičnih zavarovancev. Če se torej zavarovalnica odloči za enotno premijo $p$ za vse zavarovance, bo ta enaka uteženemu povprečju $p_1$ in $p_2$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2},$$

kar pomeni, da velja: $p_2 < p < p_1$. Predstavljajmo si, da sedaj na trg pride konkurenčna zavarovalnica, ki pa zaračunava dve različni premiji za ti dve vrsti zavarovancev, torej $p_1$ in $p_2$. Ker je $p_2 < p < p_1$, to pomeni, da se bodo manj rizični zavarovanci raje zavarovali pri konkurenčni zavarovalnici, kjer bodo plačevali premijo $p_2$, bolj rizični zavarovanci pa bodo ostali pri prvotni zavarovalnici, saj bodo

plačevali premijo $p$, ki je manjša od $p_1$. Vendar pa to pomeni, da bo zavarovalnica zaslužila $n_1 p$, potrebovala pa bi $n_1 p_1$, kar je seveda več od zaslužene premije, torej zavarovalnica s pobrano premijo ne bo mogla pokriti svojih obveznosti, kar lahko vodi v insolventnost zavarovalnice.

V zgornjem primeru smo videli, da je ključno prepoznavanje manj rizičnih in bolj rizičnih zavarovancev in posledično zaračunavanje različnih premij, saj le tako lahko zavarovalnice ostanejo konkurenčne in seveda solventne. Seveda lahko obstaja veliko faktorjev, ki vplivajo na rizičnost zavarovanca. Ti so lahko odvisni tako od lastnosti zavarovanca (starost, vozniške izkušnje, spol,...), od lastnosti zavarovančevega objekta (starost vozila, moč vozila, vrsta gradnje hiše,...) kot od lastnosti kraja, kjer se zavarovani objekt nahaja (število prebivalcev, bližina vode,...).

Nekateri od teh faktorjev so med seboj popolnoma neodvisni, lahko pa med posameznimi faktorji obstajajo tudi korelacije. Problem korelacij se pojavi, če analiziramo faktorje s pomočjo t.i. enostranske analize. V tem primeru bi namreč rizičnost zavarovanca določali zgolj na podlagi enega faktorja. Predstavljajmo si, da na rizičnost zavarovanca vplivajo samo vozniške izkušnje ne pa tudi starost avtomobila, vendar pa med tema faktorjema obstaja korelacija, saj mladi vozniki največkrat vozijo starejše avtomobile. V primeru enostranske analize bi se pri analizi faktorja vozniških izkušenj seveda pokazalo, da so mladi vozniki bolj rizični, vendar pa bi se pri enostranski analizi starosti vozila ravno tako pokazalo, da so bolj rizična starejša vozila, saj je v tem razredu prevladujoč delež mladih voznikov. To bi pomenilo, da bi lahko mladim voznikom na podlagi enostranske analize zaračunali dvakratno doplačilo kot bi si ga zaslužili, enkrat na podlagi vozniških izkušenj, in drugič zato, ker vozijo starejši avto.

Problem, ki se pojavlja pri diktiranju faktorjev, ki vplivajo na rizičnost zavarovanca, je tudi v povezavi z dostopnostjo podatkov o posameznih faktorjih. Pri nekaterih faktorjih se težko zanesemo na resničnost podatkov, saj jih je v praksi težko preverjati. Določene faktorje kot npr. spol pa je zaradi diskriminacije celo prepovedano uporabiti pri določanju cene zavarovanja.

Problem se lahko pojavi tudi, če o posameznem faktorju nimamo zgodovinskih podatkov, saj če v preteklosti nismo zbirali podatkov o tem faktorju, ne moremo na podlagi tega faktorja določiti rizičnosti zavarovanca. Dejstvo torej je, da vseh faktorjev, ki vplivajo na rizičnost zavarovanca, ne moremo uporabiti pri oblikovanju cene. Tudi če bi imeli o faktorjih dovolj podatkov, bi morali paziti, da model ne bi bil preveč kompleksen za vsakodnevno uporabo, hkrati pa bi lahko z vključitvijo prevelikega števila faktorjev povzročili prekomerno prileganje podatkom, na podlagi katerih določamo rizičnost zavarovancev, tak model pa seveda ne bi bil več uporaben. Ključno je, da vključimo faktorje, ki najbolj vplivajo na rizičnost zavarovanca, hkrati pa je podatkov o faktorju dovolj in so seveda tudi verodostojni.

## 3.1 Multiplikativni model

Ključno je, da zavarovance na podlagi njihovih lastnosti razdelimo v razrede. V istem razredu so torej zavarovanci z istimi lastnostmi oziroma istimi nivoji faktorjev, kar pomeni, da si zavarovanci znotraj istega razreda zaslužijo enako premijo. Rekli

bomo, da ti zavarovanci sodijo v isti premijski razred. Če teh faktorjev ni veliko in imamo v vsakem premijskem razredu dovolj podatkov, lahko za vsak razred izračunamo premijo kot pričakovano škodo zavarovancev v tem razredu (po zakonu velikih števil). Problem se pojavi, ko je teh faktorjev več. Predstavljajmo si, da imamo $p$ faktorjev, vsak faktor pa ima $k_i$ nivojev, $i = 1, \ldots, p$. V tem primeru je število premijskih razredov enako $\prod_{i=1}^{p} k_i$, kar pomeni, da število premijskih razredov s številom faktorjev in številom nivojev hitro narašča. Z večanjem števila razredov se lahko zgodi, da za posamezni razred nimamo dovolj podatkov, določitev premije na podlagi teh podatkov pa je nezanesljiva, saj je lahko varianca podatkov zelo velika.

Rešitev tega je, da uporabimo multiplikativni model, kar pomeni, da predpostavimo, da faktorji na rizičnost zavarovanca vplivajo multiplikativno. Metodo lahko uporabimo pri izračunu škodne pogostosti, povprečna škode ali nevarnostne premije, odvisno od tega, kaj si želimo modelirati.

Naj $i$ predstavlja $i$-ti premijski razred, $\mu_i$ pa vrednost, ki jo želimo modelirati (to je lahko škodna pogostost, povprečne škoda ali nevarnostna premija). Naj bo naš model sestavljen iz $p$ različnih faktorjev, kar pomeni, da lahko $i$-ti premijski razred zapišemo kot $i = (i_1, i_2, ..., i_p)$, kjer z $i_j$ označimo nivo j-tega faktorja, ki določa $i$-ti premijski razred. Multiplikativni model predpostavlja, da lahko $\mu_i$ zapišemo kot

$$\mu_i = \gamma_0 \gamma_{1i_1} \gamma_{2i_2} \cdots \gamma_{pi_p},$$

kjer $\gamma_{ji_j}$ za $j = 1, ..., p$ določa, koliko se rizičnost zavarovanca poveča/pomanjša na podlagi $j$-tega faktorja, če je nivo tega faktorja enak $i_j$.

Pri vseh modelih, ki jih bomo obravnavali, bomo privzeli tudi naslednje predpostavke:
Naj z $X$ označimo slučajno spremenljivko, ki označuje število škod ali skupno vsoto škod. Privzamemo, da velja:

1. Če se $X_i$ nanaša na polico $i$, kjer je $m$ število vseh polic, potem velja, da so $X_1, ..., X_m$ neodvisne slučajne spremenljivke.

2. Naj bo $m$ število disjunktnih časovnih intervalov, $X_i$ pa naj se nanaša na $i$-ti časovni interval. Potem velja, da so $X_1, ..., X_m$ neodvisne slučajne spremenljivke.

3. Naj se $X_1$ in $X_2$ nanašata na dve različni polici znotraj istega premijskega razreda. Potem velja, da imata $X_1$ in $X_2$ enako porazdelitveno funkcijo.

   ([9, str. 6, 7])

Multiplikativni model upošteva korelacije med faktorji, vendar pa ne upošteva interakcije med faktorji. Interakcija med faktorji namreč pomeni, da je določen faktor odvisen od vrednosti drugih faktorjev. Če želimo upoštevati tudi interakcijo med faktorji moramo naše podatke ustrezno transformirati, kar bomo videli v razdelku 5.13 Nadgradnja GLM.

Parametre, ki nastopajo v multiplikativnem modelu, lahko ocenimo z uporabo posplošenega linearnega modela – GLM (generalized linear model). Ker gre za posplošeni linearni model, si najprej oglejmo osnovne predpostavke linearnega modela.

# 4 Linearni model

S pomočjo linearnega modela izrazimo odvisnost spremenljivke $Y$ od neodvisnih spremenljivk $x_i$, $i = 1, \ldots, r$, ki jih zapišemo v vektor $X = [x_1, \ldots, x_r] \in \mathbb{R}^{1 \times r}$. V linearnem modelu vidimo $Y$ kot slučajno spremenljivko, izidi te slučajne spremenljivke pa so naše opazovane vrednosti, ki jih označimo z $y$. Linearni model predpostavlja, da lahko slučajno spremenljivko Y zapišemo kot vsoto njene pričakovane vrednosti $\mu$ in slučajne spremenljivke $\varepsilon$

$$Y = \mu + \varepsilon.$$

Predpostavimo, da je:

- pričakovana vrednost $Y$ enaka $\mu$, ki jo lahko zapišemo kot linearno kombinacijo spremenljivk $x_i$, $i = 1, \ldots, r$,
- napaka $\varepsilon$ normalno porazdeljena, s pričakovano vrednostjo 0 in varianco $\sigma^2$ ($\varepsilon \sim N(0, \sigma^2)$).

Skupaj z zgornjima predpostavkama lahko zgornjo enačbo zapišemo tudi drugače

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

kjer je $\mu = X\beta$, $\beta \in \mathbb{R}^r$, hkrati pa velja, da je $\varepsilon \sim N(0, \sigma^2)$, kar pomeni, da ima $Y$ ne glede na vrednosti $X$ konstantno varianco.

Zgornji linearni model lahko napišemo tudi v vektorski obliki. To pomeni, da na naše podatke gledamo kot na izid slučajnih spremenljivk $Y_1, Y_2, \ldots, Y_n$, kjer je $i$-ti podatek izid slučajne spremenljivke $Y_i$, $n$ pa število podatkov. V tem primeru je $Y$ slučajni vektor, $X$ pa matrika vrednosti neodvisnih spremenljivk, kjer se $i$-ta komponenta vektorja $Y$ nanaša na vrednosti neodvisnih spremenljivk v $i$-ti vrstici matrike X, ki jo bomo označevali z $X_i$. Če torej opazujemo $n$ izidov (podatkov), ki so odvisni od $r$ parametrov, to pomeni, da je $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times r}$, $\beta \in \mathbb{R}^r$, $\varepsilon \in \mathbb{R}^n$ in $r \leq n$, da je sistem $Y = X\beta + \varepsilon$ rešljiv. Število $r$ se nanaša na število parametrov, ki jih ocenjujemo, in je npr. v primeru, da $X$ predstavlja indikatorske spremenljivke, enako vsoti nivojev posameznih faktorjev: $r = \sum_{i=1}^{p} k_i$, kjer je $p$ število faktorjev, $k_i$ pa število nivojev $i$-tega faktorja. Indikatorske spremenljivke so spremenljivke, ki lahko zavzamejo vrednost 1 ali 0, in sicer imajo vrednost 1, če je nivo posameznega faktorja zastopan v opazovanem premijskem razredu, sicer pa 0. Za lažje razumevanje smo indikatorske spremenljivke uporabili tudi v primeru na str. 18.

**Opomba 4.1.** Zgoraj smo rekli, da se število $n$ nanaša na število izidov, v zavarovalništvu bi npr. $n$ predstavljal število polic ali pa število škodnih dogodkov. Največkrat se v praksi združuje izide, ki so v istem premijskem razredu, torej tiste izide, ki imajo enake nivoje vseh faktorjev. Vrednost $n$ v tem primeru predstavlja število premijskih razredov, $Y$ pa vrednost na ravni celotnega premijskega razreda.

Če želimo, da bo model enoličen, moramo na začetku določiti premijski razred, ki bo predstavljal t.i. izhodiščni razred (intercept), ki bo pri vsakem faktorju določil

en nivo, za katerega ne bomo ocenjevali parametra. Poglejmo si preprost primer.

Imejmo 2 faktorja, npr. spol in vozniške izkušnje, kjer obstajata dva razreda: do 3 leta vozniških izkušenj, nad 3 leta vozniških izkušenj. To pomeni, da imamo 4 premijske razrede: ženski spol – do 3 leta vozniških izkušenj, ženski spol – nad 3 leta vozniških izkušenj, moški spol – do 3 leta vozniških izkušenj, moški spol – nad 3 leta vozniških izkušenj. Naj bo matrika X sedaj sestavljena iz indikatorskih spremenljivk, kjer se 1. stolpec nanaša na ženski spol, 2. stolpec na moški spol, 3. stolpec na do 3 leta vozniških izkušenj, 4. stolpec na nad 3 leta vozniških izkušenj, vsaka vrstica pa predstavlja drug premijski razred. X bi imel torej obliko

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}.$$

V tem primeru model ni enoličen, saj če k $\beta_1$ in $\beta_2$ prištejemo neko realno število, od $\beta_3$ in $\beta_4$ pa to število odštejemo, še vedno dobimo enak vektor $Y$. Zato določimo izhodiščni razred (intercept). Naj bo v našem primeru izhodiščni razred premijski razred z nivojema: ženski spol in do 3 leta vozniških izkušenj. Zgornji model bi sedaj napisali kot

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix},$$

kjer so v prvem stolpcu vedno enke, saj se ta stolpec nanaša na parameter, ki predstavlja izhodiščni razred (intercept). Parameter $\beta_2$ se tako nanaša zgolj na razliko v učinku, ki jo moški spol povzroči na opazovano spremenljivko $Y$ v primerjavi z ženskim spolom, $\beta_3$ pa zgolj na razliko, ki jo povzročijo zavarovanci z nad 3 leti vozniškimi izkušnjami na opazovano spremenljivko $Y$ v primerjavi z zavarovanci z do 3 leti vozniških izkušenj. Tak model je sedaj enoličen.

Če želimo kar najbolj zanesljive rezultate, potem v izhodiščni razred (intercept) vzamemo tiste nivoje, ki imajo pri posameznem faktorju največjo izpostavljenost.

Predpostavke linearnega modela so torej naslednje ([1, str. 9]):

- **Slučajna komponenta:** Vsaka komponenta vektorja $Y$ je neodvisna in normalno porazdeljena slučajna spremenljivka. Pričakovana vrednost komponent je lahko različna, vse komponente pa imajo enako varianco.
- **Sistematična komponenta:** Linearna kombinacija $r$ spremenljivk nam da linearni prediktor $\eta$
$$\eta = X\beta.$$
- **Povezovalna funkcija:** Odvisnost slučajne komponente od sistematične komponente določa povezovalna funkcija. Pri linearnem modelu je povezovalna funkcija kar identiteta
$$\mathbb{E}[Y] = \mu = \eta.$$

## 4.1 Ocenjevanje parametrov

Standardni način določanja parametrov $\beta$ pri linearnem modelu je preko minimiziranja vsote kvadratov, kar pomeni, da rešujemo problem

$$F(\beta; Y, X) = \sum_{i=1}^{n} (Y_i - X_i\beta)^2,$$

$$\min_{\beta} F(\beta; Y, X),$$

kjer je $Y_i$ $i$-ta opazovana vrednost, $X_i$ pa spremenljivke, ki se nanašajo na $i$-to opazovano vrednost ($i$-ta vrstica matrike $X$).

Problem lahko zapišemo tudi kot

$$\frac{\partial}{\partial \beta_j} F(\beta; Y, X) = 0, \forall j \iff \sum_{i=1}^{n} 2(Y_i - X_i\beta)X_{ij} = 0, \forall j \iff X^T X\beta = X^T Y.$$

V nadaljevanju bomo dokazali, da je rešitev enaka

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

kjer je $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times r}$, $\hat{\beta} \in \mathbb{R}^r$.

Reševanje zgornjega problema pa je ekvivalentno metodi največjega verjetja, kjer maksimiziramo funkcijo verjetja.

Naj bo Y slučajna spremenljivka z gostoto verjetnosti $f(y; \beta)$, kjer je $\beta$ neznani parameter. Po metodi največjega verjetja je cenilka $\hat{\beta}$ za $\beta$ vrednost, ki maksimizira funkcijo verjetja $L(\beta; y) = f(y; \beta)$.

Maksimizacija funkcije verjetja je ekvivalentna maksimizaciji (naravnega) logaritma funkcije verjetja, saj je naravni logaritem strogo naraščajoča funkcija

$$l(\beta; y) = \log L(\beta; y) = \log f(y; \beta).$$

Vemo, da za linearni model velja

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

$$\mathbb{E}[Y] = X\beta.$$

To pomeni, da je opazovana vrednost $Y_i$ porazdeljena $N(X_i\beta, \sigma^2)$. Ker so po predpostavki $Y_i$ za $i = 1, 2, ..., n$ neodvisne slučajne spremenljivke, je funkcija verjetja enaka produktu gostot verjetnosti $Y_i$

$$\begin{aligned} L(\beta, \sigma^2; Y) &= \prod_{i=1}^{n} f(\beta, \sigma^2; Y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(Y_i - X_i\beta)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta)}, \end{aligned}$$

kjer smo z zapisom $L(\beta, \sigma; Y)$ in $f(\beta, \sigma; Y_i)$ poudarili, da je lahko neznan tudi parameter $\sigma^2$. Logaritem funkcije verjetja je enak

$$l(\beta, \sigma^2; y) = -\frac{n}{2}(\log(2\pi) + \log\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta).$$

Iščemo torej $\beta$ in $\sigma^2$, ki bosta maksimizirala $l(\beta, \sigma^2; y)$

$$\begin{aligned} \frac{\partial}{\partial\beta} l(\beta; y) &= \frac{1}{2\sigma^2}(Y - X\beta)^T X = 0, \\ \frac{\partial}{\partial\sigma^2} l(\beta, \sigma^2; y) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) = 0. \end{aligned}$$

Od tu dobimo

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Če je tudi parameter $\sigma^2$ neznan, potem cenilko $\hat{\sigma}^2$ za $\sigma^2$ dobimo iz enačbe

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) = 0,$$

kjer namesto $\beta$ uporabimo njeno cenilko $\hat{\beta}$. Cenilka za $\sigma^2$ je torej enaka

$$\hat{\sigma^2} = \frac{1}{n}(Y - X\hat{\beta})^T (Y - X\hat{\beta}).$$

Problem je, da je predpostavke lineranega modela v zavarovalništvu težko zadostiti. Problem povzročijo predvsem naslednji pojavi:

- Slučajna spremenljivka $Y_i$ lahko ni normalno porazdeljena. To se v zavarovalništvu pojavi, ko želimo npr. modelirati število škod. Število škod je namreč diskretna slučajna spremenljivka, hkrati pa mora biti tudi nenegativna. Problem je rešljiv, če obstaja taka transformacija $f$, da je $f(Y_i)$ normalno porazdeljena slučajna spremenljivka. Vendar pa seveda ni nujno, da taka transformacija obstaja.

- Slučajna spremenljivka $Y_i$ nima konstantne variance. Varianca $Y_i$ je na primer lahko odvisna od pričakovane vrednosti $Y_i$.

- Velikokrat $Y_i$ lahko zavzame samo nenegativne vrednosti. S predpostavko o normalni porazdelitvi in konstantni varianci tej omejitvi seveda ne moremo zadostiti, kar pomeni, da modela ne moremo uporabiti pri modeliranju $Y_i$, kjer je $Y_i \geq 0$ (npr. pri modeliranju višine škode).
- Povprečna vrednost $Y_i$ naj bi bila linearna funkcija spremenljivk $X_i$, pri oblikovanju cen v zavarovalništvu pa naj bi bil bolj primeren in uporaben že omenjeni multiplikativni model.

V zavarovalništvu in seveda tudi drugje je zato velikokrat bolj uporaben posplošen linearni model oziroma GLM, saj s posplošitvijo predpostavk linearnega modela lahko odpravimo zgornje probleme.

# 5 Posplošeni linearni model – GLM

Osnove GLM sta postavila Nelder in Wedderburn v njunem delu Generealized Linear Models leta 1972.

## 5.1 Osnovne predpostavke GLM

Kot smo omenili že zgoraj, je pri linearnem modelu v praksi težko zadostiti določenim predpostavkam. Pri GLM te predpostavke posplošimo, s tem pa je model v praksi bolj uporaben. Predpostavke GLM so naslednje ([1, str. 10]):

1. **Slučajna komponenta:** Komponente slučajnega vektorja $Y$ so neodvisne slučajne spremenljivke s porazdelitveno funkcijo iz družine eksponentno porazdeljenih slučajnih spremenljivk.
2. **Sistematična komponenta:** Linearna kombinacija $r$ spremenljivk nam da linearni prediktor $\eta$
$$\eta = X\beta.$$
3. **Povezovalna funkcija:** Odvisnost slučajne komponente od sistematične komponente določa povezovalna funkcija $g$
$$\mathbb{E}[Y] = \mu = g^{-1}(\eta).$$

Glede na zgornje predpostavke lahko vidimo, da linearni model ravno tako zadošča zgornjim predpostavkam in je zato le eden od posebnih primerov GLM.

## 5.2 Družina eksponentno porazdeljenih slučajnih spremenljivk

Porazdelitvene funkcije komponent $Y$ morajo pripadati družini eksponentno porazdeljenih slučajnih spremenljivk. To družino definira dvoparametrična porazdelitvena funkcija

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i, \phi, w_i)\right\}, \tag{1}$$

kjer je $\theta_i$ odvisen od $i$-ja, kar pomeni, da ima lahko vsak $Y_i$ različen $\theta$, $\phi > 0$ pa je neodvisen od $i$ in zato enak za vse komponente $Y_i$. Funkcija $b(\theta_i)$ se imenuje kumulativna funkcija in mora biti dvakrat zvezno odvedljiva hkrati pa mora imeti pozitiven drugi odvod, kar pomeni, da je konveksna funkcija. Predpostavili bomo tudi, da obstaja inverzna funkcija prvega odvoda funkcije $b(\theta_i)$. Z izbiro kumulativne funkcije določimo tip porazdelitve iz družine eksponentnih porazdelitev (npr. Poissonovo, normalno, gama porazdelitev), porazdelitev pa je nato enolično določena z izbiro parametrov $\theta_i$ in $\phi$.

Zgornja gostota verjetnosti je seveda definirana samo za $y_i$, ki predstavljajo možen izid $Y_i$. Za vse ostale vrednosti $y_i$ je $f_{Y_i}(y_i) = 0$. V praksi je $Y_i \in (0, \infty)$ ali pa je $Y_i \in (-\infty, \infty)$. Veljati mora tudi, da je $\phi > 0$ in $w_i \geq 0$ in da $\theta_i$ leži na odprtem intervalu, npr. $0 < \theta_i < 1$.

**Opomba 5.1.** Z utežjo $w_i$, ki jo bomo imenovali izpostavljenost, lahko posamezen $i$-ti izid ustrezno predhodno utežimo. Kaj vzamemo za izpostavljenost, je odvisno od vrednosti, ki jo želimo modelirati. V praksi je pri modeliranju škodne pogostosti izpostavljenost enaka trajanju polic, pri modeliranju povprečne škode pa številu škodnih dogodkov. V primeru, ko podatke združujemo na ravni premijskega razreda, predstavlja $w_i$ izpostavljenost $i$-tega premijskega razreda. Izpostavljenost je v GLM zelo pomemben podatek, saj se z večanjem izpostavljenosti zmanjšuje varianca modeliranih vrednosti, kar bomo videli tudi v nadaljevanju.

**Opomba 5.2.** Vrednost $\phi$ je lahko pri posameznem tipu porazdelitve, ki pripada družini eksponentnih porazdelitev, fiksna vrednost, kar pomeni, da je ta porazdelitev enoparametrična eksponentna porazdelitev (npr. Poissonova porazdelitev).

### 5.2.1 Lastnosti družine eksponentno porazdeljenih slučajnih spremenljivk

Predvsem nas bo zanimalo matematično upanje in varianca te družine slučajnih spremenljivk. Za izpeljavo teh dveh vrednosti si bomo pomagali z nekaterimi drugimi funkcijami, ki nam bodo olajšale izračune.

**Definicija 5.3.** (Začetni moment slučajne spremenljivke)
Za zvezno slučajno spremenljivko $X$ z gostoto verjetnosti $f(x)$ je začetni moment reda $r$ definiran kot

$$m_r = \int_{-\infty}^{\infty} x^r f(x) dx$$

za $r = 1, 2, ...$, za diskretno slučajno spremenljivko $X$ z verjetnostno funkcijo $\mathbb{P}(X = x_k) = p_k$ je začetni moment reda $r$ definiran kot

$$m_r = \sum_{k=0}^{\infty} x_k^r p_k$$

za $r = 1, 2, ...$

**Definicija 5.4.** (Centralni moment slučajne spremenljivke)
Za zvezno slučajno spremenljivko $X$ z gostoto verjetnosti $f(x)$ je centralni moment reda $r$ definiran kot

$$\mu_r = \int_{-\infty}^{\infty} (x - m_1)^r f(x) dx$$

za $r = 1, 2, ...$, za diskretno slučajno spremenljivko $X$ z verjetnostno funkcijo $\mathbb{P}(X = x_k) = p_k$ je centralni moment reda $r$ definiran kot

$$\mu_r = \sum_{k=0}^{\infty} (x_k - m_1)^r p_k$$

za $r = 1, 2, ...$

Možno je, da vsote oziroma integrali, s katerimi smo definirali momente, ne obstajajo, s tem pa seveda ne obstajajo pripadajoči momenti slučajne spremenljivke. Velja pa, da kadar obstaja začetni (centralni) moment reda $r$, obstaja tudi centralni (začetni) moment reda $r$ ter vsi začetni in centralni momenti nižjega reda.

Opazimo lahko

$$\mathbb{E}[X] = m_1,$$
$$\mathrm{VaR}[X] = \mu_2.$$

**Definicija 5.5.** (Momentno rodovna funkcija)
Momentno rodovna funkcija slučajne spremenljivke $X$ je definirana kot

$$M_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

Z momentno rodovno funkcijo je porazdelitev slučajne spremenljivke X enolično določena, ni pa nujno, da za vsako slučajno spremenljivko X momentno rodovna funkcija obstaja. Lastnosti momentno rodovne funkcije:

- $M_X(0) = 1$,
- $\frac{\partial^r M_X(t)}{\partial t^r}|_{t=0} = m_r$,
- Razvoj v Taylorjevo vrsto: $M_X(t) = 1 + m_1 t + \frac{m_2}{2!} t^2 + \frac{m_3}{3!} t^3 + ...$

Pomemben je tudi naravni logaritem momentno rodovne funkcije $\Psi_X(t) = \log M_X(t)$. Njegov $r$-ti odvod v točki 0 se imenuje kumulanta reda $r$

$$\kappa_r = \left. \frac{\partial^r \Psi_X(t)}{\partial t^r} \right|_{t=0}$$

Ker je $M_X(0) = 1$, je $\Psi_X(0) = 0$, razvoj v Taylorjevo vrsto okrog točke 0 pa je tako

$$\log M_X(t) = \kappa_1 t + \frac{\kappa_2}{2!} t^2 + \frac{\kappa_3}{3!} t^3 + ...$$

Hitro lahko izračunamo

$$\kappa_1 = \mathbb{E}[X], \tag{2}$$
$$\kappa_2 = \mathrm{VaR}[X]. \tag{3}$$

Zgornjo lastnost lahko sedaj uporabimo pri računanju matematičnega upanja in variance slučajne spremenljivke iz družine eksponentno porazdeljenih slučajnih spremenljivk.

Momentno rodovna funkcija slučajne spremenljivke $Y_i$ z gostoto verjetnosti (1) je

$$M_{Y_i}(t) = \int_{-\infty}^{\infty} e^{ty} f_{Y_i}(y;\theta_i,\phi)dy =$$
$$= \int_{-\infty}^{\infty} \exp\big\{\frac{y(\theta_i + t\phi/w_i) - b(\theta_i)}{\phi/w_i} + c(y,\phi,w_i)\big\}dy =$$
$$\exp\big\{\frac{b(\theta_i + t\phi/w_i) - b(\theta_i)}{\phi/w_i}\big\}\int_{-\infty}^{\infty} \exp\big\{\frac{y(\theta_i + t\phi/w_i) - b(\theta_i + t\phi/w_i)}{\phi/w_i} + c(y,\phi,w_i)\big\}dy.$$

Tu lahko uporabimo predpostavko, da vrednosti $\theta_i$ ležijo na oprtem intervalu. To pomeni, da za $t$ v okolici 0 (torej za $|t| < \delta$ za neki $\delta > 0$) vrednost $\theta_i + t\phi/w_i$ še vedno leži na istem odprtem intervalu. Zadnji integral je torej integral gostote verjetnosti slučajne spremenljivke s parametrom $\theta_i + t\phi/w_i$ po celotnem definicijskem območju, vrednost integrala pa je za $|t| < \delta$ enaka 1.
Za $|t| < \delta$ torej velja

$$M_{Y_i}(t) = \exp\left\{\frac{b(\theta_i + t\phi/w_i) - b(\theta_i)}{\phi/w_i}\right\}.$$

Iz tega sledi, da je za $|t| < \delta$ logaritem momentno rodovne funkcije enak

$$\Psi_{Y_i}(t) = \log M_{Y_i}(t) = \frac{b(\theta_i + t\phi/w_i) - b(\theta_i)}{\phi/w_i}.$$

Sedaj lahko izračunamo matematično upanje in varianco slučajne spremenljivke $Y_i$. Spomnimo se, da smo predpostavili, da je $b(.)$ dvakrat odvedljiva funkcija, torej je

$$\Psi'_{Y_i}(t) = b'(\theta_i + t\phi/w_i),$$
$$\Psi''_{Y_i}(t) = b''(\theta_i + t\phi/w_i)\phi/w_i.$$

Po (2) in (3) torej sledi

$$\mathbb{E}[Y_i] = \Psi'_{Y_i}(0) = b'(\theta_i),$$
$$\mathrm{VaR}[Y_i] = \Psi''_{Y_i}(0) = b''(\theta_i)\phi/w_i.$$

Torej $\mu_i = \mathbb{E}[Y_i] = b'(\theta_i)$. Ker vemo, da obstaja inverzna funkcija odvoda $b(.)$, velja $\theta_i = b'^{-1}(\mu_i)$, kar pa lahko vstavimo v $b''(\theta_i)$ ter dobimo t.i. funkcijo variance $v(\mu_i) = b''(b'^{-1}(\mu_i))$. Iz tega sledi, da je $\mathrm{VaR}[Y_i] = v(\mu_i)\phi/w_i$, kar pomeni, da smo varianco zapisali kot funkcijo matematičnega upanja.

Zgornje izpeljave lahko sedaj strnemo v naslednjo lemo.

**Lema 5.6.** *Naj bo $Y_i$ slučajna spremenljivka z gostoto verjetnosti* (1)*. Potem obstaja logaritem momentno rodovne funkcije $\Psi_{Y_i}(t)$, ki ga zapišemo kot*

$$\Psi_{Y_i}(t) = \frac{b(\theta_i + t\phi/w_i) - b(\theta_i)}{\phi/w_i}.$$

*Velja*

$$\mu_i = \mathbb{E}[Y_i] = b'(\theta_i),$$
$$\mathrm{VaR}[Y_i] = v(\mu_i)\phi/w_i,$$

*kjer je $v(\mu_i)$ funkcija variance, definirana z $v(\mu_i) = b''(b'^{-1}(\mu_i))$.*

Funkcija variance je zelo pomembna v GLM, kar dokazuje tudi naslednji izrek.

**Izrek 5.7.** *Znotraj množice slučajnih spremenljivk iz družine eksponentnih porazdelitev z gostoto verjetnosti* (1) *je tip porazdelitve (npr. Poissonova porazdelitev) enolično določen s funkcijo variance.*

Dokaz lahko najdemo v [5, str. 19].

To pomeni, da je pri posameznem GLM potrebno določiti le funkcijo variance in ne celotne porazdelitvene funkcije, saj je z njo enolično določen tudi tip porazdelitve.

Naslednji izrek govori o tem, da so slučajne spremenljivke z gostoto verjetnosti (1) reproduktivne.

**Izrek 5.8.** *(Reproduktivnost)*
*Naj bosta $Y_1$ in $Y_2$ neodvisni slučajni spremenljivki, ki imata isti tip porazdelitve iz družine eksponentnih porazdelitev, kar pomeni, da imata enako $b(.)$. Poleg tega zahtevamo, da imata enak $\mu$ in $\phi$, lahko pa imata različno utež $w$, npr. $w_1$ in $w_2$. Potem uteženo povprečje $Y = (w_1Y_1 + w_2Y_2)/(w_1 + w_2)$ pripada istemu tipu porazdelitve kot $Y_1$ in $Y_2$ z utežjo $w = w_1 + w_2$.*

*Dokaz.* Za logaritem momentno rodovne funkcije velja

$$\Psi_{wY}(t) = \log \mathbb{E}[e^{twY}] = \Psi_Y(wt),$$

če pa sta X in Y neodvisni slučajni spremenljivki, velja

$$\begin{aligned}\Psi_{X+Y}(t) &= \log \mathbb{E}[e^{t(X+Y)}] = \log \mathbb{E}[e^{tX}e^{tY}] = \log[\mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]] \\ &= \log \mathbb{E}[e^{tX}] + \log \mathbb{E}[e^{tY}] = \Psi_X(t) + \Psi_Y(t).\end{aligned}$$

Sedaj nas zanima momentno rodovna funkcija za $Y = (w_1Y_1 + w_2Y_2)/(w_1 + w_2)$

$$\begin{aligned}\Psi_Y(t) &= \Psi_{Y_1}(tw_1/(w_1 + w_2)) + \Psi_{Y_2}(tw_2/(w_1 + w_2)) \\ &= \frac{b(\theta + (tw_1/(w_1 + w_2))\phi/w_1) - b(\theta)}{\phi/w_1} + \frac{b(\theta + (tw_2/(w_1 + w_2))\phi/w_2) - b(\theta)}{\phi/w_2} \\ &= \frac{(w_1 + w_2)(b(\theta + t\phi/(w_1 + w_2)) - b(\theta))}{\phi} \\ &= \frac{b(\theta + t\phi/(w_1 + w_2)) - b(\theta)}{\phi/(w_1 + w_2)}.\end{aligned}$$

Ker je porazdelitev slučajne spremenljivke enolično določena z momentno rodovno funkcijo in zato tudi z logaritmom momentno rodovne funkcije, velja, da ima Y enako porazdelitev kot $Y_1$ oziroma $Y_2$, le da ima utež enako $w = w_1 + w_2$. □

Uporabnost tega izreka je v tem, da lahko v GLM v primeru, ko imata dva razreda istega faktorja isto matematično upanje, ta dva razreda združimo, saj imata po zgornjem izreku še vedno enako porazdelitev kot prej. To je z vidika oblikovanja cen v zavarovalništvu potrebno in seveda smiselno. Npr. lognormalna ali posplošena Paretova porazdelitev nimata te lastnosti in zato v primerih, ko je potrebno združevanje razredov, nista primerni za uporabo.

### 5.2.2 Porazdelitvena funkcija škodne pogostosti

Naj bo $N(t)$ število škodnih dogodkov na eno zavarovalno polico do časa $t$, kjer je $N(0) = 0$. $\{N(t); t \geq 0\}$ je slučajni proces, ki ga imenujemo škodni proces. Pokazati se da, da je pod predpostavkama, ki sta blizu 2. in 3. predpostavki, ki smo ju privzeli pri multiplikativnem modelu, in predpostavko, da se dva škodna dogodka ne moreta zgoditi hkrati, škodni proces kar Poissonov proces ([2, str. 349]). To pomeni, da je število škodnih dogodkov v nekem časovnem intervalu na eno zavarovalno polico Poissonovo porazdeljena slučajna spremenljivka. Po predpostavki GLM, ki govori o neodvisnosti zavarovalnih polic, pa je tudi vsota škodnih dogodkov po vseh zavarovalnih policah Poissonovo porazdeljena slučajna spremenljivka. Naj bo torej $Z_i$ število škodnih dogodkov za nek premijski razred $i$, ki ima izpostavljenost $w_i$, in naj bo $\mu_i$ matematično upanje, če je $w_i = 1$. Potem sledi, da je $\mathbb{E}[Z_i] = w_i\mu_i$ in je torej $Z_i$ Poissonovo porazdeljena slučajna spremenljivka z verjetnostno funkcijo

$$f_{Z_i}(z_i, \mu_i) = e^{-w_i\mu_i}\frac{(w_i\mu_i)^{z_i}}{z_i!}, \quad z_i = 0, 1, 2, ...$$

Dokaz lahko najdemo v [9, str. 8]. Ker je v praksi velikokrat bolj uporabno modeliranje škodne pogostosti kot pa število škodnih dogodkov, nas bolj zanima slučajna spremenljivka $Y_i = Z_i/w_i$, ki je ravno tako Poissonovo porazdeljena slučajna spremenljivka in jo bomo imenovali relativna Poissonova porazdelitev. Verjetnostna funkcija $Y_i$ je torej enaka

$$\begin{aligned} f_{Y_i}(y_i, \mu_i) &= \mathbb{P}(Y_i = y_i) = \mathbb{P}(Z_i = w_i y_i) \\ &= e^{-w_i\mu_i}\frac{(w_i\mu_i)^{w_i y_i}}{(w_i y_i)!} \\ &= \exp\{w_i(y_i \log \mu_i - \mu_i) + c(y_i, w_i)\} \end{aligned}$$

za tiste vrednosti $y_i$, kjer je $w_i y_i$ nenegativno celo število, saj $w_i y_i$ predstavlja število škodnih dogodkov. Funkcija $c(y_i, w_i)$ je enaka $c(y_i, w_i) = w_i y_i \log w_i - \log((w_i y_i)!)$. Zgornja verjetnostna funkcija je oblike (1), če je $\theta_i = \log \mu_i$,

$$f_{Y_i}(y_i, \mu_i) = \exp\{w_i(y_i\theta_i - e^{\theta_i}) + c(y_i, w_i)\}.$$

Vidimo lahko, da je $\phi = 1$, $b(\theta_i) = e^{\theta_i}$. Vemo, da je $\mu_i > 0$, torej je $-\infty < \theta_i < \infty$.

**Opomba 5.9.** Vprašanje je, če je predpostavka o Poissonovi porazdelitvi realistična? Torej ali so realistične predpostavke, ki so nam zagotovile Poissonov proces. Predvsem je težko zagotoviti homogenost znotraj istega premijskega razreda. Vemo, da se lahko škodna pogostost skozi leto razlikuje, vendar pa to ni problem, če opazujemo škodne dogodke skozi celo leto, saj je število škodnih dogodkov v celotnem letu tudi Poissonovo porazdeljena slučajna spremenljivka. Problem je lahko v tem, da bi morale imeti slučajne spremenljivke v istem premijskem razredu z isto izpostavljenostjo enako porazdelitveno funkcijo, kar pa ni nujno, saj je v modelu težko upoštevati prav vse faktorje, ki vplivajo na rizičnost zavarovanca. Dva zavarovanca imata torej lahko iste opazovane značilnosti, vseeno pa se bosta razlikovala v rizičnosti. To lahko ublažimo tako, da predpostavimo, da je tudi $\mu_i$ slučajna spremenljivka. V tem primeru imamo t.i. mešano Poissonovo porazdelitev, ki ima večjo varianco kot

pa standardna Poissonova porazdelitev, kar si bomo v nadaljevanju tudi še podrobneje pogledali. Taki modeli, naj bi tudi zagotovili boljše prileganje zavarovalniškim podatkom.

### 5.2.3 Porazdelitvena funkcija povprečne škode

Pri modeliranju povprečne škode nas zanima porazdelitev slučajne spremenljivke $Y_i = Z_i/w_i$, $i = 1, \ldots, r$, kjer $Z_i$ predstavlja celotno škodo v premijskem razredu $i$, $w_i$ pa število škodnih dogodkov v tem razredu. Tu je število škodnih dogodkov $w_i$ utež in ne slučajna spremenljivka, kot je bila pri modeliranju škodne pogostosti. Gre za to, da najprej modeliramo škodno pogostost in s tem dobimo število škodnih dogodkov kot izid slučajne spremenljivke, nato pa modeliramo povprečno škodo, ki jo izračunamo glede na to dobljeno število škodnih dogodkov.

Pri modeliranju škodne pogostosti smo zaradi predpostavk o homogenosti in neodvisnosti privzeli, da je število škodnih dogodkov Poissonovo porazdeljena slučajna spremenljivka. Pri modeliranju povprečne škode pa ni čisto očitno, katera porazdelitev je v danem primeru najbolj primerna. Vemo, da mora biti slučajna spremenljivka pozitivna, njena gostota verjetnosti pa asimetrična, saj so velike škode seveda manj verjetne. Zaradi tega vemo, da normalna porazdelitev ni primerna, obstaja pa veliko drugih porazdelitev, ki zadoščajo tema dvema zahtevama. V GLM modeliranju je najbolj standardna uporaba gama porazdelitve. Gre za to, da je z uporabo gama porazdelitve standardni odklon sorazmeren z $\mu_i$, kar pomeni, da je koeficient variacije CV, ki je definiran kot $CV = \frac{\sqrt{\mathrm{VaR}[Y_i]}}{\mathbb{E}[Y_i]}$, konstanten, kar pa je s praktičnega vidika smiselna predpostavka za povprečno škodo.

Predpostavimo, da je posamezna škoda gama porazdeljena slučajna spremenljivka. Ker gledamo le posamezno škodo, je $w_i = 1$. V tem primeru lahko gostoto verjetnosti gama porazdelitve napišemo kot

$$f(z) = \frac{\beta_i^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta_i z}, \quad z > 0,$$

kjer je parameter oblike $\alpha > 0$ in $\beta_i > 0$. Če bi naredili reparametrizacijo $\theta_i = 1/\beta_i$, bi parameter $\theta_i$ predstavljal parameter merila. Porazdelitveno funkcijo za gama porazdelitev bomo na kratko označevali z $\mathrm{G}(\alpha, \beta_i)$. Povprečna vrednost je enaka $\alpha/\beta_i$, varianca pa $\alpha/\beta_i^2$. Lepa lastnost gama porazdelitve je tudi ta, da je vsota neodvisnih gama porazdeljenih slučajnih spremenljivk, ki imajo isti parameter $\beta_i$, spet porazdeljena gama s parametrom $\beta_i$, parameter $\alpha$ pa je enak vsoti vseh posameznih parametrov $\alpha$. Torej, če je $Z_i$ vsota $w_i$ neodvisnih enako porazdeljenih slučajnih spremenljivk, ki so porazdeljene kot $\mathrm{G}(\alpha, \beta_i)$, potem je $Z_i \sim \mathrm{G}(w_i\alpha, \beta_i)$. Gostota verjetnosti slučajne spremenljivke $Y_i = Z_i/w_i$ je potem enaka

$$f_{Y_i}(y_i) = w_i f_Z(w_i y_i) = \frac{(w_i\beta_i)^{w_i\alpha}}{\Gamma(w_i\alpha)} y_i^{w_i\alpha-1} e^{-w_i\beta_i y_i}, \quad y_i > 0,$$

kar pomeni, da je $Y_i \sim \mathrm{G}(w_i\alpha, w_i\beta_i)$, matematično upanje pa je enako $\alpha/\beta_i$. Da bi zgornjo gostoto verjetnosti preoblikovali v obliko (1), zamenjamo $\alpha$ in $\beta_i$ z $\mu_i$ in $\phi$, kjer je $\mu_i = \alpha/\beta_i$, $\phi = 1/\alpha$. Ker je $\alpha > 0$ in $\beta_i > 0$, je tudi $\mu_i > 0$ in $\phi > 0$. Gostoto

verjetnosti $Y_i$ lahko sedaj zapišemo kot

$$f_{Y_i}(y_i) = f_{Y_i}(y_i; \mu_i, \phi) = \frac{1}{\Gamma(w_i/\phi)} \left(\frac{w_i}{\mu_i \phi}\right)^{w_i/\phi} y_i^{(w_i/\phi)-1} e^{-w_i y_i/(\mu_i \phi)}$$
$$= \exp\left\{\frac{-y_i/\mu_i - \log \mu_i}{\phi/w_i} + c(y_i, \phi, w_i)\right\}, \quad y > 0,$$

kjer je $c(y, \phi, w_i) = w_i/\phi \log(w_i y_i/\phi) - \log y_i - \log \Gamma(w_i/\phi)$. Tu je torej

$$\mathbb{E}[Y_i] = w_i \alpha/(w_i \beta_i) = \mu_i,$$

$$\mathrm{VaR}[Y_i] = w_i \alpha_i/(w_i \beta_i)^2 = \phi \mu_i^2/w_i.$$

Potrebna je še ena reparametrizacija, kjer $\mu_i$ zamenjamo s $\theta_i$, kjer je $\theta_i = -1/\mu_i$, kar pa pomeni, da je $\theta_i < 0$. Gostoto verjetnosti slučajne spremenljivke $Y_i$ je torej

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i \theta_i + \log(-\theta_i)}{a_i(\phi)} + c(y_i, \phi, w_i)\right\}.$$

Vidimo, da gama porazdelitev res spada v družino eksponentno porazdeljenih slučajnih spremenljivk, kjer je $b(\theta_i) = -\log(-\theta_i)$, tako da gamo porazdelitev lahko uporabimo v GLM modeliranju.

## 5.3 Tweediejev model

V premoženjskem zavarovalništvu si pri modeliranju želimo uporabljati porazdelitve, ki so invariantne na množenje s konstanto. To pomeni, da za porazdelitev velja, da če je $Y$ slučajna spremenljivka, katere porazdelitvena funkcija pripada določenemu tipu porazdelitve, potem tudi porazdelitvena funkcija $cY$ za $c > 0$ pripada temu tipu porazdelitve. Ta lastnost je v praksi zelo zaželena, saj je $Y$ velikokrat merjena v denarnih enotah in je smiselno, da se s spremembo denarne enote ali denarne valute tip porazdelitve slučajne spremenljivke $Y$ ne spremeni. Model za oblikovanje cen z uporabo takih porazdelitev torej ni odvisen od valute, v kateri so podani podatki, saj tip porazdelitve ostane isti, spremenijo se le vrednosti parametrov, ki nastopajo v porazdelitveni funkciji. Enako bi pri modeliranju škodne pogostosti to pomenilo, da se tip porazdelitve ne spremeni, ne glede na to, ali merimo škodno pogostost v procentih ali promilih.

Zanima nas torej, katere porazdelitve, ki so oblike (1), so tudi invariantne na množenje s konstanto. Izkaže se, da so to tiste porazdelitve, za katere velja, da se da funkcijo variance zapisati kot

$$v(\mu_i) = \mu_i^p$$

za nek $p$ ([5]). Modeli, v katerih so uporabljene take porazdelitve se imenujejo Tweediejevi modeli.

V tabeli 5.1 lahko vidimo, katere vrednosti $p$ ustrezajo kateri porazdelitvi in za kateri model je ta porazdelitev uporabna. Z izbiro $1 < p < 2$ dobimo sestavljeno Poissonovo porazdelitev ([5, str. 47]), ki je primerna za modeliranje nevarnostne premije. Slučajna spremenljivka $Y$ je porazdeljena sestavljeno Poissonovo, če je $Y = \sum_{k=1}^{N} X_k$, slučajne spremenljivke $X_1, X_2, ...$ so neodvisne in enako porazdeljene,

Tabela 5.1: Tweediejevi modeli ([9, str. 25])

| **vrednost p** | **tip porazdelitve** | **ime porazdelitve** | **modeliranje** |
|---|---|---|---|
| $p < 0$ | zvezna | – | – |
| $p = 0$ | zvezna | normalna | – |
| $0 < p < 1$ | ne obstaja | – | – |
| $p = 1$ | diskretna | Poissonova | škodne pogostosti |
| $1 < p < 2$ | mešana | sestavljena Poissonova | nevarnostne premije |
| $p = 2$ | zvezna, pozitivna | gama | povprečne škode |
| $2 < p < 3$ | zvezna, pozitivna | – | povprečne škode |
| $p = 3$ | zvezna, pozitivna | inverzna Gaussova | povprečne škode |
| $p > 3$ | zvezna, pozitivna | – | povprečne škode |

hkrati pa so neodvisne tudi od Poissonovo porazdeljene slučajne spremenljivke $N$. Tu ni potrebna ločena analiza škodne pogostosti in povprečne škode, saj če izberemo, da so $X_1, X_2, ...$ gama porazdeljene slučajne spremenljivke, slučajna spremenljivka $Y$ predstavlja celotno škodo, to pa je že dovolj za izračun potrebne nevarnostne premije. V tabeli 5.1 lahko vidimo, da je sestavljena Poissonova porazdelitev mešana porazdelitev, saj ima v točki 0 skok, pri vrednostih, ki so večje od 0, pa je porazdelitev zvezna.

Ohlsson in Johansson v [9, str. 25] navajata, da sta Jørgensen in Souza modelirala nevarnostno premijo za zavarovanje osebnih avtomobilov v Braziliji in za njune podatke je vrednost $p = 1{,}37$ maksimizirala funkcijo največjega verjetja.

Presenetljivo je, da za vrednost $0 < p < 1$ ne obstaja nobena porazdelitev oblike (1), dokaz lahko najdemo v [5, str. 46]. Za negativno vrednost $p$ sicer obstajajo zvezne porazdelitve, ampak pri modeliranju v zavarovalništvu zaenkrat še niso bile uporabljene.

Zaenkrat so nam s praktičnega vidika najbolj zanimive porazdelitve, pri katerih je $p \geq 1$. Funkcija $b(\theta)$ je za te vrednosti definirana kot

$$b(\theta) = \begin{cases} e^{\theta} & \text{za} \quad p = 1, \\ -\log(-\theta) & \text{za} \quad p = 2, \\ -\frac{1}{p-2}[-(p-1)\theta]^{(p-2)/(p-1)} & \text{za} \quad 1 < p < 2 \quad \text{in} \quad p > 2. \end{cases}$$

Če z $M_\theta$ označimo interval, kjer je parameter $\theta$ definiran, potem je

$$M_\theta = \begin{cases} -\infty < \theta < \infty & \text{za} \quad p = 1, \\ -\infty < \theta < 0 & \text{za} \quad p > 1. \end{cases}$$

Odvod $b'(\theta)$ je enak

$$\mu = b'(\theta) = \begin{cases} e^{\theta} & \text{za} \quad p = 1, \\ (-(p-1)\theta)^{-1/(p-1)} & \text{za} \quad p > 1, \end{cases}$$

inverz $b'(\theta)$, ki ga označimo s funkcijo $h$, pa je enak

$$\theta = h(\mu) = \begin{cases} \log \mu & \text{za} \quad p = 1, \\ -\frac{1}{p-1}\mu^{-(p-1)} & \text{za} \quad p > 1. \end{cases}$$

Dokaz lahko najdemo v [5, str. 43].

## 5.4 Povezovalna funkcija

V linearnem modelu smo videli, da se da matematično upanje neodvisne slučajne spremenljivke $Y_i$ zapisati kot linearno kombinacijo odvisnih spremenljivk $x_j$, $j = 1, \ldots, r$, ki so shranjene v $i$-ti vrstici matrike $X$, torej $X_i$. V posplošenem linearnem modelu to predpostavko posplošimo, saj zahtevano, da se da neko transformacijo matematičnega upanja $Y_i$ zapisati kot linearno kombinacijo spremenljivk $x_j$, $j = 1, \ldots, r$. To transformacijo pa opiše povezovalna funkcija $g(.)$, za katero mora veljati, da je monotona in odvedljiva funkcija. To pomeni, da mora za vektor $\mu = (\mu_1, \ldots, \mu_n)$ veljati

$$g(\mu) = \eta = X\beta.$$

V primeru multiplikativnega modela, ki je v zavarovalništvu zelo uporaben, je primerna logaritemska povezovalna funkcija

$$g(\mu_i) = \log \mu_i.$$

Logaritemska povezovalna funkcija je torej primerna za vse Tweediejeve modele, ki smo jih omenili v prejšnjem poglavju.

V linearnem modelu je povezovalna funkcija enaka identiteti, torej je

$$g(\mu_i) = \mu_i \quad \text{oziroma} \quad \mu_i = \eta_i.$$

V primeru, da imamo model, kjer modeliramo npr. verjetnost, uporabimo t.i. logit funkcijo

$$\eta_i = g(\mu_i) = \log\Big(\frac{\mu_i}{1-\mu_i}\Big).$$

Ta funkcija zagotovi, da bo $\mu_i \in (0, 1)$.

### 5.4.1 Kanonična povezovalna funkcija

V GLM so parametri, ki nastopajo v modelu, med sabo povezani z dvema funkcijama, $b'(.)$ in $g(.)$

$$\mu_i = b'(\theta_i),$$
$$\eta_i = g(\mu_i).$$

Funkcija $b(.)$ opiše strukturo slučajne spremenljivke in je enolično določena s funkcijo variance. Funkcija $g(.)$ pa se uporabi pri modeliranju matematičnega upanja $\mu_i$, kar pomeni, da obstaja več različnih možnosti za izbiro funkcije $g(.)$.

Kanonična povezovalna funkcija je posebna oblika povezovalne funkcije, kjer zahtevamo, da je $g(.) = b'^{-1}(.)$ kar pomeni, da je $g(\mu_i) = \eta_i = \theta_i$. Izkaže se, da uporaba kanonične povezovalne funkcije poenostavi nekatere izračune, vendar pa seveda ni nujno, da je ta funkcija tudi najbolj primerna za naše modeliranje.

Primeri:

- Linearni model: Velja, da je $\mu_i = b'(\theta_i) = \theta_i$, kar pomeni, da je identitetna povezovalna funkcija $g(\mu_i) = \mu_i$ kanonične oblike.
- Poissonov model: Za ta model velja, da je $\mu_i = b'(\theta_i) = e^{\theta_i}$, kar pomeni, da je logaritemska povezovalna funkcija $g(\mu_i) = \log \mu_i$ kanonične oblike.
- Gama model: Pri gama porazdelitvi velja, da je $\mu_i = b'(\theta_i) = -1/\theta_i$, kar pomeni, da je kanonična povezovalna funkcija $g(\mu_i) = -1/\mu_i$. Problem pri uporabi te funkcije je, da lahko vrednosti matematičnega upanja zavzamejo tudi negativne vrednosti, kar pa pri modeliranju povprečne škode ni primerno.

## 5.5 Ocenjevanje parametrov

Ko imamo postavljen model, je ključno, da z oceno parametrov ta model čimbolj približamo danim podatkom. Želimo torej dobiti vektor $\beta$, da bo veljalo

$$g(\mu) = X\beta.$$

Problem rešujemo z metodo največjega verjetja. Najprej pa si poglejmo še nekaj lastnosti metode največjega verjetja.

Naj bo $X$ slučajna spremenljivka z gostoto verjetnosti $f(x;\theta)$. Funkcija verjetja $L(\theta;x)$ je enaka

$$L(\theta;x) = f(x;\theta).$$

Maksimizacija funkcije verjetja je ekvivalentna maksimizaciji logaritma funkcije verjetja

$$l(\theta;x) = \log L(\theta;x) = \log f(x;\theta).$$

**Definicija 5.10.** (Fisherjeva informacija)
Fisherjeva informacija slučajne spremenljivke $X$ z gostoto verjetnosti $f(x;\theta)$ je definirana kot

$$I = \mathbb{E}[(l'(\theta;X))^2] = \int_{-\infty}^{\infty} [l'(\theta;x)]^2 f(x;\theta)dx.$$

S pomočjo Fisherjeve informacije merimo, koliko informacij nosi o neznanem parametru $\theta$ slučajna spremenljivka $X$, katere porazdelitev je odvisna od parametra $\theta$.

Fisherjevo informacijo pa lahko izračunamo tudi na drugačen način.

**Izrek 5.11.** *Fisherjeva informacija slučajne spremenljivke $X$ z gostoto verjetnosti $f(x;\theta)$ je enaka*

$$I = -\mathbb{E}[l''(\theta;X)].$$

*Dokaz.* Prvi odvod logaritma funkcije verjetja po $\theta$ je enak

$$l'(\theta;x) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta) = \frac{1}{f(x;\theta)}f'(x;\theta),$$

kjer s črtico $'$ označujemo odvod po $\theta$. Drugi odvod pa

$$l''(\theta;x) = \frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{\partial}{\partial\theta}\left(\frac{1}{f(x;\theta)}f'(x;\theta)\right) =$$

$$= \frac{f''(x;\theta)f(x;\theta) - [f'(x;\theta)]^2}{f^2(x;\theta)} = \frac{f''(x;\theta)}{f(x;\theta)} - [l'(\theta;x)]^2.$$

Iz tega sledi, da je

$$\mathbb{E}[l''(\theta;x)] = \int_{-\infty}^{\infty} \Big[\frac{f''(x;\theta)}{f(x;\theta)} - [l'(\theta;x)]^2\Big] f(x;\theta)dx =$$

$$= \int_{-\infty}^{\infty} f''(x;\theta)dx - \mathbb{E}[(l'(\theta;x))^2] = \int_{-\infty}^{\infty} f''(x;\theta)dx - I.$$

Integral dvakratnega odvoda gostote verjetnosti pa je enak 0, saj velja

$$\int_{-\infty}^{\infty} f''(x;\theta)dx = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial\theta^2} f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2} \int_{-\infty}^{\infty} f(x;\theta)dx = \frac{\partial^2}{\partial\theta^2} 1 = 0.$$

Torej je

$$\mathbb{E}[l''(\theta;x)] = -I.$$

□

Kakšna bi bila lahko neformalna interpretacija Fisherjeve informacije?
Fisherjevo informacijo se izračuna kot $I = \mathbb{E}[(l'(\theta;X))^2]$, hkrati pa vemo, da je $l'(\theta;x) = \frac{f'(x;\theta)}{f(x;\theta)}$. Vidimo lahko, da je Fisherjeva informacija slučajne spremenljivke z gostoto verjetnosti $f(x;\theta)$ večja, če se $f(x;\theta)$ s spreminjanjem vrednosti $\theta$ hitrejše spreminja. V tem primeru bo že majhna sprememba v $\theta$ zelo spremenila obliko gostote verjetnosti, kar pomeni, da bomo v tem primeru iz podatkov lažje natančnejše ocenili, katera porazdelitev se najbolj prilega podatkom, saj se porazdelitve s spreminjanjem parametra $\theta$ med seboj zelo razlikujejo. Če pa je Fisherjeva informacija majhna, to pomeni, da se $f(x;\theta)$ s spreminjanjem $\theta$ ne spreminja hitro in so porazdelitve s podobnim parametrom $\theta$ med seboj podobne in jih je zato težje med seboj razlikovati, s tem pa tudi hitreje pride do napak pri iskanju porazdelitve, ki se najbolj prilegajo podatkom.

Naj bodo sedaj $X_1, X_2, ..., X_n$ neodvisne in enako porazdeljene slučajne spremenljivke z gostoto verjetnosti $f(x;\theta_i)$, $i = 1, ..., n$. Zaradi neodvisnosti je njihova skupna gostota verjetnosti enaka produktu posameznih gostot verjetnosti, temu pa je enaka tudi funkcija verjetja

$$L_n(\theta;x) = f_n(x;\theta) = \prod_{i=1}^{n} f(x_i;\theta_i),$$

kjer je $x \in \mathbb{R}^n$ in $\theta \in \mathbb{R}^r$. Po metodi največjega verjetja torej maksimiziramo funkcijo verjetja $L_n(\theta;x)$, kar je ekvivalentno maksimizaciji logaritma funkcije verjetja

$$l_n(\theta;x) = \log L_n(\theta;x) = \log\left[\prod_{i=1}^{n} f(x_i;\theta_i)\right] = \sum_{i=1}^{n} \log f(x_i;\theta_i) = \sum_{i=1}^{n} l(\theta_i;x_i).$$

Ker pa vemo, da iščemo vektor $\beta$ in da je $\mu_i = b'(\theta_i)$ in $g(\mu_i) = \eta_i = X_i\beta$, kjer je $X_i$ $i$-ta vrstica matrike $X$, potem vemo, da je $\theta_i$ odvisna od $\beta$. V tem primeru je Fisherjeva informacijska matrika enaka

$$I = \mathbb{E}[(l_n'(\theta(\beta);x))^2] = -\mathbb{E}[l''(\theta(\beta);x)],$$

dokaz pa je podoben kot pri eni dimenziji.

Ko je parameter $\beta \in \mathbb{R}^r$ (kot je v našem primeru), je Fisherjeva informacijska matrika $I \in \mathbb{R}^{r \text{ x } r}$. Po definiciji pa lahko Fisherjevo informacijsko matriko $I$ zapišemo tudi kot

$$I = -\mathbb{E}[H(l(\theta(\beta); x))],$$

kjer je $H(l(\theta(\beta); x))$ Hessejeva matrika funkcije $l(\theta(\beta); x)$, ki jo bomo označevali kar s $H$. Hessejeva matrika funkcije $l(\theta(\beta); x)$ je namreč kvadratna matrika, $H \in \mathbb{R}^{r \text{ x } r}$, ki jo sestavljajo drugi parcialni odvodi funkcije $l(\theta(\beta); x)$ po $\beta$.

Sedaj se lahko vrnemo k oceni parametrov $\beta$. Vemo, da so slučajne spremenljivke $Y_1, Y_2, ..., Y_n$ po predpostavki neodvisne in enako porazdeljene slučajne spremenljivke z gostoto verjetnosti oblike (1). Funkcijo verjetja $L(\theta; \phi, y)$ lahko zapišemo kot produkt gostot verjetnosti slučajnih spremenljivko $Y_i$ za $i = 1, ..., n$, logaritem funkcije verjetja pa označimo z $l(\theta; \phi, y)$. Logaritem funkcije verjetja je torej enak

$$l(\theta; \phi, y) = \frac{1}{\phi} \sum_{i=1}^{n} w_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi, w_i).$$

Z metodo največjega verjetja želimo dobiti zgolj oceno za $\theta$, medtem ko si bomo oceno za parameter $\phi$ ogledali v naslednjem poglavju.Vrednost $\theta$, ki bo maksimizirala logaritem funkcije verjetja, ne bo odvisna od predznaka vrednosti $\phi$, saj je po predpostavki $\phi > 0$. To pa pomeni, da pri določanju ocene parametrov $\theta$ ne potrebujemo vrednosti $\phi$.

Funkcijo smo sedaj zapisali kot funkcijo $\theta$. Ker vemo, da nas zanima ocena za vektor $\beta$, si želimo logaritem funkcije verjetja $l(\theta; \phi, y)$ zapisati kot funkcijo $\beta$. Vemo, da velja $\mu_i = b'(\theta_i)$ in $g(\mu_i) = \eta_i = X_i\beta$, potem vemo, da je $\theta_i$ odvisna od $\beta$ in lahko funkcijo verjetja maksimiziramo glede na vrednost $\beta$ tako, da funkcijo verjetja $l(\beta; \phi, y)$ odvajamo glede na $\beta$ s pomočjo posrednega odvoda preko $\theta$

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} w_i(y_i - b'(\theta_i)) \frac{\partial \theta_i}{\partial \beta_j}$$

$$= \frac{1}{\phi} \sum_{i=1}^{n} w_i(y_i - b'(\theta_i)) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

Izračunamo lahko

$$\mu_i = b'(\theta_i) \Rightarrow \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = v(\mu_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = 1/b''(\theta_i) = 1/v(\mu_i),$$

kjer smo upoštevali, da je odvod inverzne funkcije enak inverzni funkciji odvoda in da je $b''(\theta_i) = v(\mu_i)$. Analogno izračunamo tudi

$$\eta_i = g(\mu_i) \Rightarrow \frac{\partial \eta_i}{\partial \mu_i} = g'(\mu_i) \Rightarrow \frac{\partial \mu_i}{\partial \eta_i} = 1/g'(\mu_i),$$

$$\eta_i = X_i\beta \Rightarrow \frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

kjer je $x_{ij}$ $(i,j)$-ti element matrike $X$. Dobljene izračune lahko vstavimo v zgornjo enačbo in dobimo

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)} x_{ij} \quad \text{za} \quad j = 1, ..., r,$$

funkcijo pa imenujemo funkcija prileganja (score function).

Pri iskanju maksimuma zgornje enačbe seveda izenačimo z 0, hkrati pa jih lahko še pomnožimo s $\phi$

$$\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)} x_{ij} = 0 \quad \text{za} \quad j = 1, ..., r. \tag{4}$$

Pozabiti ne smemo, da je $\mu_i = \mu_i(\beta)$ in da mora rešitev zgornje enačbe torej zadostiti tudi naslednjim pogojem

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(X_i\beta). \tag{5}$$

To pomeni, da je rešitev$\mu_i = y_i$, ki reši enačbo (4), primerna samo v primeru, ko je število parametrov enako številu opazovanj (torej $r = n$), saj bomo le v tem primeru lahko zadostili tudi enačbi (5).

Enačbo (4) lahko zapišemo v matrični obliki, če definiramo nove uteži, ki so odvisne tudi od parametrov

$$\tilde{w_i} = \frac{w_i}{v(\mu_i)g'(\mu_i)},$$

diagonalna matrika $W$ pa naj ima na diagonali elemente $\tilde{w_i}$, $i = 1, ..., n$. Enačba (4) je tako enaka

$$X^T W y = X^T W \mu,$$

kjer je $y \in \mathbb{R}^{n\text{x}1}$, $\mu \in \mathbb{R}^{n\text{x}1}$, $X \in \mathbb{R}^{n\text{x}r}$, $W \in \mathbb{R}^{n\text{x}n}$.

V primeru linearnega modela je $\tilde{w_i} = w_i$, če pa vzamemo še primer, ko je $w_i = 1$, dobimo standardno enačbo za linearno regresijo $X^T y = X^T X\beta$, saj velja $\mu = X\beta$, matrika $W$ pa je v tem primeru identična matrika.

Enačbo $X^T W y = X^T W \mu$ se skoraj vedno rešuje numerično.

## 5.6 Modeliranje nevarnostne premije

Za določanje cen v zavarovalništvu potrebujemo nevarnostno premijo. Pri GLM modeliranju se lahko tega lotimo na dva načina. Prvi način je seveda z uporabo Tweediejevega modela, kjer je $1 < p < 2$, drugi način pa je z modeliranjem tako škodne pogostosti kot povprečne škode, nevarnostna premija pa je v tem primeru zmnožek teh dveh vrednosti. Razlogi, da se v praksi večkrat odločimo za ločeno obravnavo škodne pogostosti in povprečne škode, so:

- Škodna pogostost je običajno bolj stabilna kot povprečna škoda, hkrati pa na škodno pogostost običajno vpliva več faktorjev kot pa na povprečno škodo. To pomeni, da so lahko faktorji tako bolj natančno ocenjeni.
- Z ločeno analizo dobimo boljši vpogled na to, kako posamezen faktor vpliva na nevarnostno premijo.

## 5.7 Oblikovanje GLM

Preden se lotimo izračuna modela po metodi največjega verjetja, se moramo odločiti, katere faktorje in kakšne razrede bomo uporabili v našem modelu. Želimo si namreč, da bi v model vključili kar največ faktorjev, ki vplivajo na določitev rizičnosti zavarovanca in s tem na višino premije.

Prednost uporabe GLM pri določanju cen je, da s pomočjo standardnih statističnih metod ocenimo, kako dobro se izbrani model prilega našim podatkom in kako bi lahko ta model izboljšali.

### 5.7.1 Testiranje hipotez in ocenjevanje $\phi$

Naj $l(\hat{\mu})$ označuje logaritem funkcije verjetja. Vemo, da če je število parametrov enako številu opazovanj ($r = n$), potem dobimo popolno prileganje, če izberemo, da je $\hat{\mu}_i = y_i$. Takemu modelu pravimo nasičeni model (saturated model). Seveda ta model v praksi ni uporaben, vseeno pa nam zaradi popolnega prileganja predstavlja neko osnovo pri določanju, kako dobro se naš model prilega podatkom. To mero imenujemo skalirana devianca (scaled deviance) $D^*$, ki predstavlja statistiko za test prileganja (likelihood-ratio-test statistic) našega modela proti modelu, ki predstavlja popolno prileganje ($n = r$). To pomeni, da testiramo hipotezo

$$H_0 : \mu = \mu_{poln\,model}$$
$$H_1 : \mu = \mu_{naš\,model}$$

Statistiko za test prileganja torej izračunamo kot

$$D^* = D^*(y, \hat{\mu}) = 2[l(y) - l(\hat{\mu})],$$

kjer smo predpostavili, da je $\phi$ v obeh modelih enak. Če velja ničelna hipoteza, potem je $D^* \sim \chi^2_{n-r}$.

Naj $h$ označuje inverzno funkcijo $b'(.)$, kar pomeni, da iz $\mu_i = b'(\theta_i)$ dobimo $\theta_i = h(\mu_i)$. To pomeni, da lahko zgornjo enačbo za $D^*$ skupaj z upoštevanjem enačbe

$$l(\theta; \phi, y) = \frac{1}{\phi} \sum_{i=1}^{n} w_i(y_i\theta_i - b(\theta_i)) + \sum_{i=1}^{n} c(y_i, \phi, w_i)$$

zapišemo kot

$$D^* = \frac{2}{\phi} \sum_{i=1}^{n} w_i(y_i h(y_i) - b(h(y_i)) - y_i h(\hat{\mu}_i) + b(h(\hat{\mu}_i))).$$

Če skalirano devianco pomnožimo s $\phi$, dobimo devianco $D = \phi D^*$.

Devianco $D$ lahko interpretiramo kot uteženo vsoto razlike med ocenjenim matematičnim upanjem $\hat{\mu}_i$ in opazovanimi vrednostmi $y_i$. Če definiramo

$$d(y, \mu) = 2[yh(y) - b(h(y)) - yh(\mu) + b(h(\mu))],$$

potem lahko devianco zapišemo tudi kot

$$D = \sum_{i=1}^{n} w_i d(y_i, \mu_i).$$

Če $d(y, \mu_i)$ opazujemo kot funkcijo $\mu_i$ pri nekem fiksnem $y_i$, lahko vidimo, da je za $\mu_i < y_i$ ta funkcija padajoča, za $\mu_i > y_i$ pa naraščajoča, v točki $y_i$ pa je $d(y_i, y_i) = 0$, saj je

$$\frac{\partial}{\partial \mu_i} d(y, \mu_i) = 2[-yh'(\mu_i) + b'(h(\mu_i))h'(\mu_i)] = 2h'(\mu)(\mu - y).$$

Po predpostavki je $b''$ pozitivna funkcija, ker pa je $h = b'^{-1}$, iz tega sledi, da je $h' = 1/b''(b'^{-1})$, torej je tudi $h'$ pozitivna funkcija.

Vidimo lahko tudi, da je maksimiziranje funkcije verjetja ekvivalentno minimiziranju deviance, če je parameter $\phi$ konstanten.

## 5.8 Pearsonov $\chi^2$-test in ocenjevanje parametra $\phi$

Pearsonov $\chi^2$-test je test, ki se uporablja za merjenje, kako dobro se podatki prilegajo izbranemu modelu. Statistika za ta test je definirana kot

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\mathrm{VaR}[Y_i]},$$

kjer so torej $y_i$ opazovane vrednosti, $\hat{\mu}_i$ pa ocenjene vrednosti. V GLM lahko torej $\chi^2$ izračunamo kot

$$\chi^2 = \frac{1}{\phi} \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

Vidimo lahko, da je v primeru linearne regresije $\chi^2 = D^*$. V Poissonovem primeru, kjer je $\phi = 1$ in $v(\mu) = \mu$, pa je $\chi^2$ klasični Pearsonov $\chi^2$

$$\chi^2 = \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

kjer je velikokrat $w_i = 1$. Pearsonov $\chi^2$ naj bi bil približno porazdeljen kot porazdelitev $\chi^2$ z $(n - r)$ prostostnimi stopnjami, kjer je $n$ število opazovanj, $r$ pa število ocenjenih parametrov. Za porazdelitev $\chi^2$ z $(n - r)$ prostostnimi stopnjami velja, da je $\mathbb{E}[\chi^2] = n - r$. To pa nam pomaga pri določitvi nepristranske cenilke za $\phi$, ki jo označimo s $\hat{\phi}_\chi$. Za nepristransko cenilko mora veljati, da je

$$\mathbb{E}[\hat{\phi}_\chi] = \phi,$$

kar je izpolnjeno, če je $\hat{\phi}_\chi$ enaka

$$\hat{\phi}_\chi = \frac{\phi\chi^2}{n - r} = \frac{1}{n - r} \sum_i w_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

Podobno lahko cenilko izračunamo tudi s pomočjo deviance

$$\hat{\phi}_D = \frac{\phi D^*(y, \hat{\mu})}{n - r} = \frac{D(y, \hat{\mu})}{n - r}$$

ali pa določimo cenilko za $\phi$ po metodi največjega verjetja.

**Opomba 5.12.** Do sedaj smo z $y_i$ označevali neko povprečno vrednost podatkov, ki padejo v $i$-ti premijski razred, preko $w_i$ pa smo v model dobili podatek o izpostavljenosti (to je lahko podatek o trajanju police, o številu škod,...) teh polic, $y_i = 1/w_i \sum_k y_{ik}$, kjer je $y_{ik}$ vrednost $k$-te individualne police v $i$-ti tarifni celici. Podatke smo torej agregirali glede na posamezne premijske razrede, kar nima vpliva na določanje parametrov $\beta$, saj bi dobili enak rezultat tudi, če podatkov ne bi agregirali. Pri določanju cenilk $\phi$ pa je priporočljivo, da podatkov ne agregiramo, saj z agregiranjem izgubimo določene informacije.

### 5.8.1 Postavljanje GLM modela

Ko začnemo s postavljanjem modela, je ključno, da najprej določimo, katere faktorje bomo vključili v model. Pomembno je seveda, da vključimo tiste faktorje, ki vplivajo na rizičnost zavarovanca. Statistična pomoč pri odločitvi, ali naj faktor vključimo v model, je testiranje hipotez. Ena možnost je s pomočjo testa verjetja (LRT), ki primerja model, ki ima vključen določen faktor, ki nas zanima, z modelom, ki tega faktorja nima vključenega.

**Lema 5.13.** *Naj bosta $H_r$ in $H_s$ dva modela, za katera velja, da je $H_s \subset H_r$. Naj bo $\hat{\mu}^{(r)}$ ocena $\mu$ po metodi največjega verjetja za model $H_r$ in $\hat{\mu}^{(s)}$ za model $H_s$. Statistika za test verjetja, kjer testiramo $H_s$ proti $H_r$, je enaka $D^*(y, \hat{\mu}^{(s)}) - D^*(y, \hat{\mu}^{(r)})$.*

Gre torej za testiranje hipoteze:

$$H_0 : \mu = \hat{\mu}^{(s)}$$
$$H_1 : \mu = \hat{\mu}^{(r)}.$$

Zgornja lema sledi iz tega, da je statistika za test verjetja določenega modela proti polnemu modelu enaka $D^* = 2[l(y) - l(\hat{\mu})]$, torej je statistika za test verjetja $H_s$ modela proti $H_r$ modelu enaka $2[l(\hat{\mu}^{(r)}) - l(\hat{\mu}^{(s)})]$, kar pa je seveda enako $D^*(y, \hat{\mu}^{(s)}) - D^*(y, \hat{\mu}^{(r)})$. To pomeni, da je pod ničelno hipotezo $D^*(y, \hat{\mu}^{(s)}) - D^*(y, \hat{\mu}^{(r)}) \sim \chi^2_{p_r - p_s}$, kjer je $p_r$ število parametrov modela $H_r$, $p_s$ pa število parametrov modela $H_s$.
Če bi namesto $\phi$ uporabili cenilko $\hat{\phi}_\chi$, potem statistiko za test verjetja preoblikujemo v

$$\frac{\phi}{\hat{\phi}_\chi}(D^*(y, \hat{\mu}^{(s)}) - D^*(y, \hat{\mu}^{(r)})) = \frac{D(y, \hat{\mu}^{(s)}) - D(y, \hat{\mu}^{(r)})}{\hat{\phi}_\chi}.$$

Cenilka $\hat{\phi}_\chi$ je izračunana iz večjega modela, torej $H_r$, saj bi v primeru, da bi cenilko izračunali iz manjšega modela, cenilka nosila tudi morebitno variacijo, ki je posledica vključitve dodatnega faktorja v model.

Pri določanju, kateri faktorji statistično značilno vplivajo na rizičnost zavarovanca, si največkrat pomagamo s $p$-vrednostjo, ki je funkcija vzorca. Ta nam pove, kolikšna je

verjetnost, da je ničelna hipoteza pravilna, mi pa jo bomo zavrnili. Če je $p$-vrednost torej manjša od vnaprej izbrane stopnje tveganja $\alpha$, potem podatki ne zadoščajo ničelni hipotezi, kar pomeni, da bomo ničelno hipotezo zavrnili.

S tem testom lahko ugotovimo, ali faktor statistično značilno vpliva na rizičnost zavarovanca ali ne. V primeru, da imamo v modelu že vključene faktorje, ki so z opazovanim faktorjem korelirani, bomo s tem testom posredno testirali tudi to, ali ta faktor prinaša v model kakšne dodatne informacije o rizičnosti zavarovanca. Problem pa je, da nam ta test nič ne pove o tem, koliko razredov tega faktorja je statistično značilnih. Faktor je namreč lahko statistično značilen tudi, če so zgolj zavarovanci znotraj enega razreda tega faktorja bolj ali manj rizični kot zavarovanci v vseh drugih razredih. Pri določanju, kateri razred statistično značilno vpliva na rizičnost zavarovanca, pa si pomagamo z intervali zaupanja ocenjenih parametrov $\hat{\beta}$.

## 5.9 Interval zaupanja na podlagi Fisherjeve informacije

S pomočjo metode največjega verjetja smo ocenili parameter $\hat{\beta}$. Seveda pa je ta cenilka kot vrednost, ki smo jo izračunali na podlagi vzorca, lahko zelo nezanesljiva. Zaradi tega določimo interval $(L, D)$, ki ga imenujemo interval zaupanja, v katerem bo z neko stopnjo zaupanja ležal parameter $\beta$. Stopnja zaupanja je največkrat določena kot vrednost $1-\alpha$, kjer je $\alpha$ stopnja tveganja in je izbrana vnaprej. Stopnja zaupanja nam torej pove, s kolikšno verjetnostjo bo parameter $\beta$ pripadal intervalu $(L, D)$. Večja je stopnja zaupanja, širši bo interval zaupanja.

### 5.9.1 Fisherjeva informacija

Fisherjeva informacija $I \in \mathbb{R}^{r \text{ x } r}$ je definirana kot negativno matematično upanje Hessejeve matrike $H \in \mathbb{R}^{r \text{ x } r}$, ki je sestavljena iz drugih odvodov logaritma funkcije verjetja $l$. Matrika H ima torej na $(j,k)$-tem mestu vrednost

$$\begin{aligned}\frac{\partial^2 l}{\partial\beta_j\partial\beta_k} &= \sum_i \frac{w_i}{\phi}\frac{\partial}{\partial\mu_i}\left[\frac{y_i-\mu_i}{v(\mu_i)g'(\mu_i)}\right]x_{ij}\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k}\\ &= \sum_i \frac{w_i}{\phi}\frac{\partial}{\partial\mu_i}\left[\frac{y_i-\mu_i}{v(\mu_i)g'(\mu_i)}\right]x_{ij}\frac{1}{g'(\mu_i)}x_{ik}\\ &= -\sum_i x_{ij}a_i x_{ik},\end{aligned}$$

kjer je

$$a_i = \frac{w_i}{\phi v(\mu_i)g'(\mu_i)^2}\left(1+(y_i-\mu_i)\frac{[v(\mu_i)g''(\mu_i)+v'(\mu_i)g'(\mu_i)]}{v(\mu_i)g'(\mu_i)}\right), \quad i=1,...,n.$$

Matriko H lahko zapišemo tudi kot $H = -X^T AX$, kjer je $A \in \mathbb{R}^{n \text{ x } n}$ diagonalna matrika, ki ima na diagonali elemente $a_i$. Ker je $\mathbb{E}[Y_i] = \mu_i$, lahko matematično upanje matrike A zapišemo z diagonalno matriko $D$, ki ima na diagonali vrednosti

$$d_i = \frac{w_i}{\phi v(\mu_i)g'(\mu_i)^2}.$$

Fisherjeva informacija se torej izračuna kot

$$I = -\mathbb{E}[H] = X^T \mathbb{E}[A] X = X^T D X.$$

Za Tweediejev model z logaritemsko povezovalno funkcijo se to poenostavi v

$$d_i = \frac{w_i}{\phi \mu_i^{p-2}},$$

kar pomeni, da Fisherjeva informacija raste linearno z $w_i$.

### 5.9.2 Interval zaupanja

Po teoriji je cenilka $\hat{\beta}$, ki jo izračunamo po metodi največjega verjetja, porazdeljena (asimptotično) normalno z matematičnim upanjem $\beta$, kovariančna matrika pa je enaka inverzni matriki Fisherjeve informacije $I$

$$\hat{\beta} \sim N(\beta, I^{-1}).$$

To pa pomeni, da je slučajna spremenljivka $I(\hat{\beta} - \beta)$ porazdeljena standardizirano normalno. Standardizirana normalna porazdelitev je simetrična, kar pomeni, da bo tudi interval zaupanja simetričen glede na 0, torej lahko interval zaupanja definiramo kot $(-z, z)$, $z \in \mathbb{R}^r$. Zanima nas torej vektor $z$, ki ga bomo izračunali po komponentah, in sicer iščemo $z_i$ za $i = 1, \ldots, r$, da bo veljalo

$$\mathbb{P}(-z_i < \sqrt{i_i}(\hat{\beta}_i - \beta_i) < z_i) = 1 - \alpha_i,$$

kjer je $i_i$ diagonalni element matrike $I$, $\alpha_i$ pa neka izbrana stopnja tveganja. Točko $z_i$ lahko dobimo iz statističnih tabel za standardizirano normalno porazdelitev, saj je $z_i = \Phi^{-1}(1 - \alpha_i/2)$, kjer je $\Phi$ porazdelitvena funkcija standardizirano normalno porazdeljene slučajne spremenljivke. To pomeni, da so v primeru, ko je $\alpha_i = \alpha$ za vsak $i = 1, \ldots, r$, komponente vektorja $z$ enake, kar bomo predpostavili tudi v nadaljnjih izračunih. Pri 95 % stopnji zaupanja je torej $z_i = 1,96$. S preoblikovanjem zgornje enačbe dobimo

$$\mathbb{P}\left(\hat{\beta}_i - \frac{z_i}{\sqrt{i_i}} < \beta_i < \hat{\beta}_i + \frac{z_i}{\sqrt{i_i}}\right) = 1 - \alpha,$$

torej je interval zaupanja za parameter $\beta_i$ enak $\left(\hat{\beta}_i - \frac{z_i}{\sqrt{i_i}}, \hat{\beta}_i + \frac{z_i}{\sqrt{i_i}}\right)$.

V primeru, da je povezovalna funkcija $g$ enaka logaritemski funkciji, lahko namesto vektorja $\beta$ gledamo vektor $\gamma \in \mathbb{R}^r$, ki ga po komponentah definiramo kot

$$\gamma_j = g^{-1}(\beta_j) = e^{\beta_j}.$$

Ker vemo, da velja

$$\mu_i = e^{X_i \beta_i}$$

za $i = 1, \ldots, n$, lahko v primeru, ko matrika $X$ predstavlja indikatorske spremenljivke, $\mu_i$ zapišemo kot

$$\mu_i = X_i \gamma.$$

V tem primeru je interval zaupanja s stopnjo zaupanja $1-\alpha$ za parameter $\gamma_i$ enak

$$\left(e^{\hat{\beta}_i - \frac{z_i}{\sqrt{i_i}}}, e^{\hat{\beta}_i + \frac{z_i}{\sqrt{i_i}}}\right).$$

Vemo torej, da bo parameter $\gamma_i$ z verjetnostjo $1-\alpha$ ležal znotraj tega intervala. V primeru, da interval zaupanja $\gamma_i$ ne vsebuje točke 0, vemo, da je verjetnost, da se razred $i$ razlikuje po rizičnosti od izhodiščnega razreda, enak $1-\alpha$. Če bo torej celoten interval zaupanja ležal levo od 0, to pomeni, da je razred $i$ manj rizičen od izhodiščnega razreda, če pa bo celoten interval zaupanja ležal desno od 0, pa je razred $i$ seveda bolj rizičen.

V primeru, da interval zaupanja za $\gamma_i$ vsebuje točko 0, pa ne moremo z verjetnostjo $1-\alpha$ trditi, da se razred $i$ razlikuje od izhodiščnega razreda, kar pomeni, da ta razred lahko priključimo izhodiščnemu razredu (interceptu). Seveda moramo po združitvi razredov ponovno oceniti tako parametre $\beta_i$ oziroma $\gamma_i$ kot tudi intervale zaupanja za te parametre.

## 5.10 Numerično iskanje rešitve po metodi največjega verjetja

Zgoraj smo pokazali, kako eksplicitno poiščemo cenilko $\hat{\beta}$ po metodi največjega verjetja. V zavarovalništvu imamo lahko več sto tisoč podatkov, ki jih modeliramo, in v tem primeru je eksplicitno iskanje cenilke po metodi največjega verjetja nepraktično. V praksi se pri iskanju cenilk zato raje poslužujemo numeričnih metod.

### 5.10.1 Newton-Raphsonova metoda

Newton-Raphsonova ali tangentna metoda je numerična metoda, ki se uporablja za iskanje ničel funkcije

$$f(x) = 0.$$

Naj bo $x^{[0]}$ neki začetni približek. Potem je iteracijska formula, po kateri iz približka $x^{[n]}$ izračunamo naslednji približek $x^{[n+1]}$, enaka

$$x^{[n+1]} = x^{[n]} - \frac{f(x^{[n]})}{f'(x^{[n]})}.$$

Metodo lahko posplošimo tudi za večdimenzionalni problem in jo uporabimo tudi pri iskanju cenilke $\hat{\beta}$ po metodi največjega verjetja, in sicer po iteracijski formuli

$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} - (H^{-1})^{[n]}(X^T W^{[n]} y - X^T W^{[n]} \mu^{[n]}),$$

kjer je $H$ Hessejeva matrika funkcije $l(\theta; X)$. Vidimo lahko, da moramo $H$, $W$ in $\mu$ preračunati na vsakem koraku iteracije.

Če je začetni približek $\hat{\beta}^{[0]}$ blizu $\hat{\beta}$, potem Newton-Raphsonova metoda hitro konvergira k $\hat{\beta}$. Seveda je lahko izračun Hessejeve matrike $H$ zelo zahteven, zato lahko metodo poenostavimo tako, da naredimo nov izračun $H$ samo na primer vsakih 10 iteracij. Problem Newton-Raphsonove metode je tudi v tem, da je lahko zelo nestabilna. Če je začetni približek $\hat{\beta}^{[0]}$ slabo izbran in se torej ne nahaja blizu $\hat{\beta}$, potem

lahko metoda celo ne konvergira. To lahko včasih odpravimo z uporabo iteracijske formule, kjer delamo manjše korake

$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} - \gamma(H^{-1})^{[n]}(X^T W^{[n]} y - X^T W^{[n]} \mu^{[n]}),$$

kjer je $\gamma$ konstanta, $0 < \gamma < 1$. Alternativni ali dodatni način stabiliziranja metode je z uporabo iteracijske formule

$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} - \gamma(H + S^{[n]} S^{[n]T})^{-1} S^{[n]},$$

kjer je $S^{[n]} = X^T W^{[n]} y - X^T W^{[n]} \mu^{[n]}$. Ta iteracijska formula še zmanjša korak, ko je vrednost $S^{[n]}$ velika.

Razlogov za nestabilnost metode je več, eden od njih je ta, da je Hessejeva matrika $H$ lahko negativno definitna, če začetni približek $\hat{\beta}^{[0]}$ ni blizu $\hat{\beta}$. Razlog lahko leži tudi v tem, da je matrika $H$ singularna, torej neobrnljiva matrika. V takih primerih lahko namesto Newton-Raphsonove metode uporabimo Fisherjevo metodo.

### 5.10.2 Fisherjeva metoda

Fisherjeva metoda je numerična metoda, pri kateri Hessejevo matriko zamenjamo z matematičnim upanjem Hessejeve matrike in tako dobimo iteracijsko formulo

$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} + (I^{-1})^{[n]}(X^T W^{[n]} y - X^T W^{[n]} \mu^{[n]}).$$

Matriko $I$ je velikokrat lažje izračunati, hkrati pa je tudi pozitivno definitna, če model ni prekomerno parametriziran. To pomaga k stabilizaciji metode, vendar pa je včasih tudi tu potrebna iteracijska formula

$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} + \gamma(I + S^{[n]} S^{[n]T})^{-1} S^{[n]} \quad \text{ali}$$
$$\hat{\beta}^{[n+1]} = \hat{\beta}^{[n]} + \gamma(I + S^{[n]T} S^{[n]} E)^{-1} S^{[n]},$$

kjer je $E$ identična matrika. Z uporabo kanonične oblike povezovalne funkcije je Fisherjeva metoda ekvivalentna Newton-Raphsonovi metodi, saj je $I = -H$.

*Dokaz.* Za kanonično povezovalno funkcijo velja

$$g'(\mu_i) = \frac{1}{v(\mu_i)},$$

$$g''(\mu_i) = -\frac{v'(\mu_i)}{v^2(\mu_i)}.$$

Diagonalna matrika A z diagonalnimi elementi $a_i$ se torej poenostavi v

$$\begin{aligned} a_i &= \frac{w_i}{\phi v(\mu_i) g'(\mu_i)^2} \left(1 + (y_i - \mu_i)\frac{v(\mu_i)g''(\mu_i) + v'(\mu_i)g'(\mu_i)}{v(\mu_i)g'(\mu_i)}\right) \\ &= \frac{w_i}{\phi v(\mu_i) g'(\mu_i)^2} \left(1 + (y_i - \mu_i)\frac{-v(\mu_i)\frac{v'(\mu_i)}{v^2(\mu_i)} + v'(\mu_i)\frac{1}{v(\mu_i)}}{v(\mu_i)\frac{1}{v(\mu_i)}}\right) \\ &= \frac{w_i}{\phi v(\mu_i) g'(\mu_i)^2}. \end{aligned}$$

Iz tega pa sledi, da je $A = D$, torej je $I = -H$. □

Glede na to, da je Fisherjeva metoda načeloma bolj stabilna, Newton-Raphsonova pa z uporabo začetnega približka $\hat{\beta}^{[0]}$ blizu $\hat{\beta}$ hitrejša, je najboljši numerični način za iskanje cenilke $\hat{\beta}$ tak, da na začetku uporabimo stabilnejšo Fisherjevo metodo, ko pa se po določenem številu iteracijskih korakov približamo vrednosti $\hat{\beta}$, uporabimo hitrejšo Newton-Raphsonovo metodo.

## 5.11 Reziduali

Pomemben člen tako pri linearni regresiji kot pri GLM je validacija izbranega modela. Pomembno je, da preverimo, kako se izbrani model prilega našim podatkom. Reziduali nam povedo, kakšna je razlika med opazovanimi vrednostmi in vrednostmi, ki smo jih dobili s pomočjo modela (fitted value). Tako nam reziduali povedo, do kakšnih razlik lahko pride, saj vemo, da nekaterih razlik v rizičnosti zaradi naključnosti ne moremo opisati z modelom.

V linearnem modelu smo reziduale lahko izračunali s preprosto formulo

$$r_i = y_i - \hat{\mu}_i.$$

Ta formula pri GLM ni več uporabna, saj vemo, da je varianca slučajne spremenljivke odvisna od njenenga matematičnega upanja. Npr. pri Poissonovem modelu bi to pomenilo, da bi z uporabo zgornje formule tudi varianca rezidualov rastla sorazmerno z $\hat{\mu}_i$. Zaradi tega je priporočljivo, da pri GLM reziduale standardiziramo na tak način, da bodo imeli reziduali v primeru, ko se izbrani model dobro prilega našim podatkom, približno enako varianco ne glede na vrednost $\hat{\mu}_i$.

### 5.11.1 Pearsonovi reziduali

Najbolj očiten način za standardizacijo je, da reziduale delimo s količino, ki je sorazmerna s standardnim odklonom. Takim rezidualom pravimo Pearsonovi reziduali

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/w_i}},$$

Pearsonovi pa se imenujejo zato, ker je vsota teh rezidualov enaka neskaliranemu Pearsonovemu $\chi^2$, $\sum_i^n (r_i^P)^2 = \phi\chi^2$.

V primeru pravilno izbranega modela bi bilo matematično upanje teh rezidualov približno enako 0, varianca pa bi bila približno enaka $\phi$.

Poznamo tudi še standardizirane Pearsonove reziduale, ki se izračunajo kot

$$r_i^{SP} = \frac{y_i - \hat{\mu}_i}{\sqrt{\phi(1-h_i)v(\hat{\mu}_i)/w_i}},$$

kjer smo jih torej standardizirali s $\phi$, s $h_i$ pa popravimo reziduale zaradi tega, ker je tudi vrednost $\mu_i$ zgolj ocenjena vrednost. Vrednost $h_i$ je diagonalni element matrike

$$H = D^{\frac{1}{2}} X (X^T D X)^{-1} X^T D^{\frac{1}{2}}.$$

Vrednost $h_i$ nam pove, v kolikšni meri je ocenjena vrednosti $\hat{\mu}_i$ odvisna od opazovanih vrednosti. Vrednost se vedno nahaja med 0 in 1. Če je vrednost blizu vrednosti

1, to pomeni, da se bo v primeru, ko opazovano vrednost malo spremenimo, tudi modelirana vrednost $\hat{\mu}_i$ malo spremenila. Rezidual se bo v tem primeru malo spremenil, saj je modelirana vrednost v veliki meri odvisna od opazovanih vrednosti. Ko torej Pearsonov rezidual delimo z $1 - h_i$, rezidual ni več odvisen od povezave med modelirano vrednostjo in opazovano vrednostjo.

### 5.11.2 Reziduali deviance

V praksi so lahko Pearsonovi reziduali nesimetrični okoli vrednosti 0. Reziduali deviance so zato v tem pogledu bolj primerni. Reziduali deviance nam tako kot pri linearnem modelu povejo, v kolikšni meri opazovana vrednost $i$ pripomore k celotni devianci.

Reziduale deviance torej izračunamo kot

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{w_i d(y_i, \hat{\mu}_i)},$$

kjer je $d(y_i, \hat{\mu}_i)$ razdalja, ki smo jo definirali pri devianci na str. 35.

Vidimo lahko, da je $\sum_i^n (r_i^D)^2 = \phi D^*$. Ko enačbo pomnožimo s predznakom razlike $y_i - \hat{\mu}_i$, dobimo tudi podatek o tem, ali je opazovana vrednost večja ali manjša od modelirane vrednosti.

Tudi tu lahko reziduale variance še standardiziramo

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{\frac{w_i d(y_i, \hat{\mu}_i)}{\phi(1 - h_i)}}.$$

### 5.11.3 Grafi rezidualov

Ko reziduale izračunamo, je ključno, da vrednosti rezidualov narišemo na graf, saj se iz grafa lažje razbere, ali pride do kakšnih nepričakovanih odstopanj. S pomočjo grafov lahko razberemo, kako dobro se izbrani model prilega našim podatkom.

Obstaja več različnih možnosti za risanje grafov rezidualov

- **Graf standardiziranih rezidualov v odvisnosti od modeliranih vrednosti $\hat{\mu}_i$:** Če so na grafu vidni trendi v zvezi z matematičnim upanjem rezidualov, to pomeni, da je v izbranem modelu nekaj narobe s predpostavko o matematičnem upanju odvisne spremenljivke $Y_i$. Mogoče v modelu nismo modelirali vseh odvisnosti med faktorji ali pa smo izbrali napačno povezovalno funkcijo. Če pa opazimo trend v zvezi v varianco rezidualov, pa to pomeni, da je bila izbrana napačna funkcija, ki opisuje odvisnost variance od matematičnega upanja. V tem primeru smo torej za $Y_i$ najverjetneje izbrali napačno porazdelitveno funkcijo.

- **Graf standardiziranih rezidualov v odvisnosti od vseh potencialnih faktorjev, ki bi lahko vplivali na $Y_i$**: Če bi v tem primeru opazili kakšen trend v zvezi z matematičnim upanjem rezidualov, bi to lahko pomenilo, da kakšnega faktorja, ki očitno vpliva na $Y_i$, nismo vključili v model.

- **Normalni Q-Q graf:** Q-Q graf je graf, kjer primerjamo dve različni porazdelitvi, in sicer primerjamo kvantile dveh porazdelitev. Če sta si porazdelitvi podobni, bodo vse točke približno ležale na premici $y = x$. Pri modeliranju podatkov uporabljamo Q-Q graf zato, da preverimo prileganje podatkov neki teoretični porazdelitvi. Pri normalnem Q-Q grafu torej preverjamo, ali so podatki približno normalno porazdeljeni. Ker vemo, da je npr. Poissonova porazdelitev za dovolj velike vrednosti $\lambda$ podobna normalni porazdelitvi, lahko s pomočjo tega grafa vidimo, ali je izbira Poissonove porazdelitve za $Y_i$ primerna za naše podatke. $Y_i$ bi tu moral predstavljati število škod in ne škodne pogostosti.

- **Grafi standardiziranih rezidualov v odvisnosti od $\mathbf{h_i}$:** S pomočjo tega grafa lahko vidimo, kateri podatki imajo močnejši vpliv na modelirane vrednosti. Če bi kakšen podatek močno odstopal od povprečja, kar se tiče velikega vpliva na modelirane vrednosti, potem bi bilo mogoče primernejše, če bi ta podatek izključili iz modeliranja.

## 5.12 Prekomerna varianca (overdispersion)

Pri GLM se lahko zgodi, da je varianca podatkov večja, kot pa je pričakovana varianca porazdelitve, ki smo jo uporabili v modelu. To v našem primeru namreč pomeni, da je varianca znotraj posameznega premijskega razreda večja, kot pa jo npr. predpostavlja izbrana Poissonova porazdelitev.

Razlogov, da pride do tega, je več, lahko pa jih načeloma ločimo v dve skupini:

- **Napačno izbrane predpostavke**: To pomeni, da na primer v model nismo vključili vseh faktorjev, ki vplivajo na rizičnost zavarovanca ali pa smo uporabili napačno povezovalno funkcijo ali napačno porazdelitveno funkcijo $Y_i$. Te napake bi lahko odkrili s pomočjo zgoraj opisanih analiz rezidualov in ustrezno spremenili model.

- **Prekomerna varianca:** Če so predpostavke v modelu pravilne, pa je varianca podatkov vseeno večja, kot pa je pričakovana varianca porazdelitve, ki smo jo uporabili v modelu.

V GLM lahko pride do prekomerne variance, ker smo privzeli, da je varianca slučajne spremenljivke $Y_i$ odvisna od matematičnega upanja. V linearnem modelu do prekomerne variance ne pride, saj je varianca $\sigma^2$ ocenjena neodvisno od matematičnega upanja $\mu$.

Razlog za to, da analiziramo prekomerno varianco, je v tem, da smo v primeru, da v našem modelu obstaja prekomerna varianca, varianco $Y_i$ s postavljenim modelom podcenili, s tem pa smo podcenili tudi intervale zaupanja ocenjenih parametrov $\beta_i$. Lahko se tudi zgodi, da bomo zato, da bi zmanjšali varianco, v model vključili preveč faktorjev, kar bo povzročilo, da se bo naš model prekomerno prilegal (overfitting) podatkom in bo v praksi tudi zelo kompleksen.

### 5.12.1 Modeliranje prekomerne variance

Ko ugotovimo, da v našem modelu pride do prekomerne variance, moramo osnovni model ustrezno razširiti in se s tem bolj približati podatkom, ki jih modeliramo. Za modeliranje prekomerne variance obstaja več različnih modelov. V večini primerov gre za enega od dveh načinov modeliranja:

- Prekomerno varianco modeliramo s funkcijo variance.
- Prekomerno varianco modeliramo tako, da parametre v porazdelitveni funkciji slučajne spremenljivke $Y_i$, ki v osnovnem modelu nastopajo kot konstante, jemljemo kot slučajne spremenljivke z neko določeno porazdelitveno funkcijo.

Če prekomerno varianco modeliramo s funkcijo variance, to pomeni, da bi v izračun variance dodali nek nov parameter, s katerim bi modelirali prekomerno varianco. Eden od načinov je, da varianco $Y_i$ sedaj zapišemo kot

$$\mathrm{VaR}[Y_i] = (1+\nu)\phi\frac{v(\mu_i)}{w_i},$$

kjer smo varianco slučajne spremenljivke $Y_i$ povečali za konstanten faktor $\nu$. Tu smo torej privzeli, da prekomerna varianca ni odvisna ne od izpostavljenosti $w_i$ kot tudi ne od matematičnega upanja $\mu_i$. Obstajajo tudi modeli, ki predpostavljajo, da faktor $\nu$ ni konstanten in je na primer odvisen od $w_i$ ali $\mu_i$. V tem primeru moramo v modelu oceniti parametre $\beta$ in pa tudi dodaten parameter $\nu$. Za ta namen se namesto standardne metode največjega verjetja uporabi t.i. kvazi metodo največjega verjetja.

Prekomerno varianco pa lahko modeliramo tudi tako, da parametre, ki nastopajo v porazdelitveni funkciji $Y_i$, jemljemo kot slučajne spremenljivke. V primeru Poissonove porazdelitve bi to pomenilo, da so $Y_i$ porazdeljeni Poissonovo s parametrom $\Lambda_i$, $Y_i|\Lambda_i \sim \mathrm{Pois}(\Lambda_i)$, kjer je $\Lambda_i$ slučajna spremenljivka, ki ima $\mathbb{E}[\Lambda_i] = \mu_i/w_i$ in $\mathrm{VaR}[\Lambda_i] = \sigma_i^2$. Iz tega sledi, da je

$$\mathbb{E}[Y_i|\Lambda_i] = \Lambda_i \quad \text{in}$$

$$\mathrm{VaR}[Y_i|\Lambda_i] = \Lambda_i.$$

Iz tega lahko s pomočjo formul

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|\Lambda_i]],$$

$$\mathrm{VaR}[Y_i] = \mathbb{E}[\mathrm{VaR}[Y_i|\Lambda_i]] + \mathrm{VaR}[\mathbb{E}[Y_i|\Lambda_i]]$$

izračunamo tudi $\mathbb{E}[Y_i]$ in $\mathrm{VaR}[Y_i]$

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i|\Lambda_i]] = \mathbb{E}[\Lambda_i] = \mu_i/w_i \quad \text{in}$$

$$\mathrm{VaR}[Y_i] = \mathbb{E}[\mathrm{VaR}[Y_i|\Lambda_i]] + \mathrm{VaR}[\mathbb{E}[Y_i|\Lambda_i]] = \mathbb{E}[\Lambda_i] + \mathrm{VaR}[\Lambda_i] = \mu_i/w_i + \sigma_i^2.$$

Vidimo, da smo torej varianco slučajne spremenljivke $Y_i$ povečali za varianco $\Lambda_i$. V primeru, ko prekomerna varianca ni odvisna od $i$, bi varianco $\Lambda_i$ preprosto definirali

kot $\mathrm{VaR}[\Lambda_i] = \sigma^2$.

V zgornjem primeru nismo določili, kakšno porazdelitveno funkcijo ima $\Lambda_i$, pač pa samo njeno matematično upanje in varianco, kar pa ni dovolj, da določimo obliko porazdelitvene funkcije za $Y_i$. Zato se za slučajno spremenljivko $\Lambda_i$ največkrat predpostavi, da je $\Lambda_i \sim \mathrm{G}(\alpha, \beta_i)$, kar pomeni

$$\mathbb{P}(Y_i = y_i | \Lambda_i = \lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad \text{in}$$

$$f_{\Lambda_i}(\lambda_i) = \frac{\beta_i^\alpha}{\Gamma(\alpha)}\lambda_i^{\alpha-1}e^{-\beta_i\lambda_i},\ \lambda_i > 0.$$

Iz tega pa lahko izračunamo porazdelitveno funkcijo $Y_i$

$$\begin{aligned}
\mathbb{P}(Y_i = y_i) &= \int_0^\infty \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\frac{\beta_i^\alpha}{\Gamma(\alpha)}\lambda_i^{\alpha-1}e^{-\beta_i\lambda_i}d\lambda_i \\
&= \frac{\beta_i^\alpha}{y_i!\Gamma(\alpha)}\int_0^\infty e^{-(1+\beta_i)\lambda_i}\lambda_i^{y_i+\alpha i-1}d\lambda_i \\
&= \frac{\beta_i^\alpha}{y_i!\Gamma(\alpha)}\frac{\Gamma(y_i+\alpha)}{(\beta_i+1)^{y_i+\alpha}}\int_0^\infty \frac{(\beta_i+1)^{y_i+\alpha}}{\Gamma(y_i+\alpha)}e^{-(1+\beta_i)\lambda_i}\lambda_i^{y_i+\alpha i-1}d\lambda_i \\
&= \frac{(y_i+\alpha-1)!}{y_i!(\alpha-1)!}\frac{\beta_i^\alpha}{(\beta_i+1)^{y_i+\alpha}} \\
&= \binom{y_i+\alpha-1}{y_i}\Big(\frac{\beta_i}{1+\beta_i}\Big)^\alpha\Big(\frac{1}{1+\beta_i}\Big)^{y_i}.
\end{aligned}$$

To pomeni, da je $Y_i$ porazdeljena negativno binomsko s parametroma $\alpha$ in $\frac{1}{1+\beta_i}$, $Y_i \sim \mathrm{NB}(\alpha, \frac{1}{1+\beta_i})$. Iz tega sledi

$$\mathbb{E}[Y_i] = \mathbb{E}[\Lambda_i] = \frac{\alpha}{\beta_i},$$

$$\mathrm{VaR}[Y_i] = \mathbb{E}[\Lambda_i] + \mathrm{VaR}[\Lambda_i] = \frac{\Big(\frac{1}{1+\beta_i}\Big)\alpha}{\Big(1+\frac{1}{1+\beta_i}\Big)^2} = \frac{(1+\beta_i)\alpha}{\beta_i^2}.$$

Že pri Tweediejevem modelu smo videli, da lahko varianco $\Lambda_i \sim \mathrm{G}(\alpha, \beta_i)$ zapišemo kot funkcijo matematičnega upanja

$$\mathrm{VaR}[\Lambda_i] = \frac{1}{\beta_i}\mathbb{E}[\Lambda_i],$$

kar pa pomeni, da velja predpostavka, da je varianca $Y_i$ odvisna od njenega matematičnega upanja, saj je

$$\mathrm{VaR}[Y_i] = \mathbb{E}[\Lambda_i] + \mathrm{VaR}[\Lambda_i] = \mathbb{E}[\Lambda_i] + \frac{1}{\beta_i}\mathbb{E}[\Lambda_i] = \Big(1+\frac{1}{\beta_i}\Big)\mathbb{E}[Y_i].$$

Iz tega lahko torej vidimo, da če lahko varianco $\Lambda_i$ zapišemo kot funkcijo njenega matematičnega upanja, potem bo tudi varianca $Y_i$ funkcija matematičnega upanja, ne glede na to, kakšno porazdelitev ima $\Lambda_i$.

## 5.13 Nadgradnja GLM

### 5.13.1 Interakcija med faktorji

Multiplikativni model predpostavlja, da so faktorji med seboj neodvisni. V praksi pa se lahko zgodi, da obstaja interakcija med faktorji, kar pomeni, da je določen faktor odvisen od drugih faktorjev, to pa moramo seveda upoštevati tudi v GLM.

V zavarovalništvu je interakcija velikokrat vidna med faktorjema spol in starost zavarovanca. Praviloma velja, da so mladi fantje (npr. starosti 18–25 let) veliko bolj rizični kot pa mlada dekleta iste starosti, medtem ko za zavarovance v srednjih letih spol naj ne bi imel več velikega vpliva na rizičnost zavarovanca. To pomeni, da pri interakciji med faktorji ne velja multiplikativnost, saj je kršeno pravilo neodvisnosti, ker se vpliv spola na rizičnost zavarovanca spreminja glede na starost zavarovanca.

Interakcijo med faktorji lahko v model vnesemo preko definiranja novega faktorja, ki je sestavljen iz faktorjev, ki so odvisni med sabo (v zgornjem primeru bi bil to faktor spol-starost). Število razredov tega faktorja je enako produktu števil razredov faktorjev, ki so med seboj odvisni.

Seveda je ključno, da najprej preverimo, med katerimi faktorji bi lahko obstajala interakcija. Interakcijo med dvema faktorjema lahko opazimo s pomočjo grafa. Če modeliramo škodno pogostost, bi torej en graf predstavljal logaritem škodne pogostosti določenega razreda prvega faktorja v odvisnosti od razredov drugega faktorja, kar pomeni, da bi bilo število grafov enako številu razredov prvega faktorja, na $x$ osi pa bi bilo toliko točk, kot je število razredov drugega faktorja (priporočljivo je, da torej faktor z večjim številom razredov damo na $x$ os). Logaritem škodne pogostosti vzamemo zato, da relativno razliko med faktorji lahko opazujemo kot absolutno razliko. Če bodo grafi med seboj vzporedni, to pomeni, da med faktorjema ni interakcije, saj bo drugi faktor ne glede na vrednosti prvega faktorja relativno enako vplival na rizičnost zavarovanca. V primeru, da grafi ne bodo vzporedni, pa pomeni, da obstaja interakcija med faktorjema, ki jo je potrebno upoštevati pri postavljanju GLM. S pomočjo grafa tudi vidimo, katere razrede lahko med seboj združimo in tako dobimo kar najmanjše število parametrov, ki jih moramo ocenjevati v modelu.

### 5.13.2 Večnivojski faktorji

V zgornjem modelu smo predpostavljali, da imajo vsi faktorji neko omejeno število razredov. To pomeni, da je imel faktor že na začetku le nekaj nivojev (npr. faktor spol) ali pa smo nivoje določenega faktorja preprosto združili v naprej določene razrede (npr. faktor vozniške izkušnje bi lahko razvrstili v 4 razrede, manj kot tri leta/4–8 let/9–15 let/nad 15 let). Faktor vozniške izkušnje je torej zvezna spremenljivka, ki ima neskončno število nivojev, vendar smo jih na podlagi linearne urejenosti razvrstili v določene razrede.

Razvrstitev v razrede je velikokrat ključna, ker lahko obstaja veliko nivojev, ki so med podatki slabo zastopani, kar pomeni, da bi bila tudi ocena parametrov $\hat{\beta}_i$ za te nivoje lahko zelo nezanesljiva. Združitev teh nivojev z drugimi nivoji bi torej lahko poskrbela za bolj zanesljive ocene parametrov $\hat{\beta}_i$. Hkrati imajo lahko nekateri nivoji tudi zelo podoben vpliv na rizičnost zavarovanca, kar pomeni, da nam združitev teh nivojev ravno tako poda bolj zanesljivo oceno parametrov, saj je ocena boljša, če

imamo večjo izpostavljenost, hkrati pa s tem zmanjšamo število parametrov, ki jih ocenjujemo. Seveda moramo paziti, da so ti razredi kar se da homogeni. Pri združevanju si lahko pomagamo tudi s podatkom o statistični značilnosti posameznega razreda. V primeru, da imata dva razreda podobno oceno parametra $\hat{\beta}_i$, lahko razmislimo o združitvi teh dveh razredov. V primeru, da pa nek razred ni statistično značilen, pa je mogoče potrebna drugačna razvrstitev.

Obstajajo pa tudi faktorji, ki imajo veliko število nivojev, vendar pa ti nivoji niso urejeni in jih je zato težko razvrstiti v razrede. Primer takega faktorja je npr. model avta. Podoben faktor je tudi poštna številka. Tu bi lahko združevali nivoje glede na to, kateri kraji so blizu drug drugega, vendar pa se v praksi velikokrat pokaže, da ni nujno, da kraji, ki so blizu drug drugega, tudi podobno vplivajo na rizičnost zavarovanca, ampak je pomembno tudi to, kakšna je gostota poselitve v tem kraju, razvitost kraja in podobno.

Problem teh faktorjev je torej v tem, da ima faktor sicer neko končno število nivojev, vendar je to število tako veliko, da je zastopanost nekaterih nivojev med podatki lahko zelo majhna in zato za te nivoje ne moremo zanesljivo oceniti $\hat{\beta}_i$. Hkrati je nivoje težko razvrstiti v določene razrede, saj nivoji niso urejeni. Tak faktor imenujemo večnivojski faktor, pri katerem ne moremo uporabiti zgoraj opisanega GLM pri izračunu cenilk $\hat{\beta}_i$. V primeru večnivojskega faktorja bi si lahko pomagali npr. s teorijo kredibilnosti.

Videli smo, da lahko GLM na različne načine nadgradimo skupaj z drugimi teorijami, ki se uporabljajo pri oblikovanju cen, in s tem prilagodimo našim podatkom. V delu smo predstavili osnovni GLM in nakazali nekatere možne nadgradnje, dejstvo pa je, da je možnih nadgradenj še zelo veliko.

# 6 Primer

Za lažje razumevanje si bomo uporabo GLM pogledali še na preprostem primeru. Pri tem si bomo pomagali s programskim orodjem R, ki se uporablja predvsem za statistično analizo podatkov. Podatke smo pridobili iz knjižnice *insuranceData*, ki so shranjeni v tabeli z imenom *dataCar*. Podatki vključujejo 67.856 avtomobilskih polic, ki so bile sklenjene v letih 2004 in 2005, 4.624 teh polic pa je imelo škodo. Za vsako polico obstajajo naslednji podatki (imena spremenljivk bomo ohranili taka kot so v tabeli *dataCar*):

- *veh_value* – vrednost vozila v 10.000 $
- *exposure* – trajanje police, ki je izraženo kot delež enega leta,
- *clm* – pove, ali je bila na polici škoda (1 – Da, 2 – Ne),
- *numclaims* – število škodnih dogodkov na polici,
- *claimcst0* – višina škode (če ni bilo škode, je vrednost 0),
- *veh_body* – tip vozila (razredi so: SEDAN, HBACK, STNWG, UTE,...),

- *veh_ age* – starost vozila (nivoji so že združeni v razrede: $1, 2, 3, 4$, kjer razred 1 predstavlja najmlajša vozila),
- *gender* – spol,
- *area* – območje (možni razredi so: A, B, C, D, E, F),
- *age_ cat* – starost voznika (nivoji so že združeni v razrede: $1, 2, 3, 4, 5, 6$, kjer razred 1 predstavlja najmlajše voznike).

Tu lahko vidimo, da so pri nekaterih spremenljivkah podatki že združeni v razrede. Nivoje združujemo v razrede zato, ker imamo lahko spremenljivke z velikim številom nivojev ali celo neskončno mnogo v primeru zveznih spremenljivk, z združevanjem nivojev v razrede pa poskrbimo za bolj pregleden model, pri katerem ocenjujemo manjše število parametrov, in pa seveda za večjo statistično značilnost. Pri tem moramo paziti, da nivoje združujemo v čimbolj homogene razrede, kar pomeni, da so v istem razredu tisti nivoji, ki podobno vplivajo na rizičnost zavarovanca. Najboljše je, da nivoje najprej združimo s pomočjo nekega logičnega reda, nato pa si pomagamo še s testiranjem hipotez, kar bomo videli tudi v nadaljevanju.

V našem primeru smo se odločili za ločeno modeliranje škodne pogostosti in povprečne škode.

## 6.1 Modeliranje škodne pogostosti

Ko imamo podatke urejene, nas seveda zanima, katere faktorje vključimo v model za škodno pogostost. Za modeliranje škodne pogostosti smo izbrali Poissonovo porazdelitev z logaritemsko povezovalno funkcijo, kjer smo za utež $w$ vzeli trajanje polic (*exposure*). V programu R smo s testom verjetja za vsak faktor primerjali model, ki ima vključen ta faktor, z modelom, ki tega faktorja nima vključenega. V prvem koraku smo v model vključili kar vseh 5 faktorjev, kar pomeni, da smo s testom za vsak faktor preverili ali je boljše, da je faktor vključen v model ali ne (ali prinese kakšne dodatne informacije o rizičnosti zavarovanca) ob predpostavki, da so v modelu vsi ostali 4 faktorji. Pri testiranju hipotez smo si izbrali stopnjo tveganja $\alpha = 5\ \%$. Pri tem smo dobili naslednje rezultate

Tabela 6.1: Model za škodno pogostost – vključenih vseh 5 faktorjev

| **Faktor** | $\hat{\chi}^2$ | **Df** | ***p*-vrednost** |
|---|---|---|---|
| ***veh_ body*** | 42,80 | 12 | 0,00 % |
| ***veh_ age*** | 30,13 | 3 | 0,00 % |
| ***gender*** | 0,61 | 1 | 43,50 % |
| ***area*** | 11,01 | 5 | 5,12 % |
| ***agecat*** | 86,07 | 5 | 0,00 % |

V tabeli stolpec $\hat{\chi}^2$ predstavlja statistiko za test verjetja, stolpec Df predstavlja število prostostnih stopenj (degrees of freedom), kar je ravno število nivojev posameznega faktorja, zmanjšano za ena (za nivo v izhodiščnem razredu), v zadnjem

stolpcu pa vidimo izračunano $p$-vrednost. Poudarili bi, da je $p$-vrednost tu zaokrožena na stotino procenta, kar pomeni, da vrednost $0,00\,\%$ pomeni, da je $p$-vrednost strogo manjša od $0,005\,\%$. Ker smo si izbrali $\alpha = 5\,\%$, to pomeni, da faktorja *gender* in *area* nista statistično značilna, saj je $p$-vrednost večja od stopnje tveganja. To pomeni, da so v modelu tudi faktorji, katerih vpliv na rizičnost zavarovanca ni statistično značilen. Naš naslednji korak je, da iz modela izključimo faktorje, ki niso statistično značilni. Faktorje je iz modela potrebno izločati po korakih, saj so faktorji med seboj lahko korelirani in z izločitvijo enega faktorja vplivamo na statistično značilnost drugih faktorjev. S poskušanjem lahko ugotovimo, da moramo iz modela odstraniti oba faktorja, *gender* in *area*, da dobimo model, v katerem so vsi faktorji statistično značilni ($p$-vrednost je manjša od $5\,\%$), rezultati pa so predstavljeni v tabeli 6.2.

Tabela 6.2: Model za škodno pogostost – vključeni statitično značilni faktorji

| **Faktor** | $\hat{\chi}^2$ | **Df** | ***p*-vrednost** |
|---|---|---|---|
| ***veh_body*** | 43,09 | 12 | 0,00 % |
| ***veh_age*** | 30,70 | 3 | 0,00 % |
| ***agecat*** | 90,73 | 5 | 0,00 % |

Poudariti je potrebno, da lahko napačno združevanje nivojev v razrede povzroči, da je faktor statistično neznačilen. V tem primeru lahko najprej razmislimo, če smo mogoče naredili napako pri združevanju nivojev in smo v isti razred združili preveč nehomogene nivoje, kar bi lahko povzročilo statistično neznačilnost tega faktorja.

Ko imamo izbrane faktorje, ki jih bomo vključili v model, lahko ocenimo parametre. Rezultati so zbrani v tabeli 6.3.

Glede na $p$-vrednost smo v tabeli 6.3 določili stopnje značilnosti, kjer smo v stolpcu značilnost uporabili naslednje oznake:

- kjer je $0\,\% \leq p$-vrednost $\leq 0,1\,\%$, je oznaka enaka ***,
- kjer je $0,1\,\% < p$-vrednost $\leq 1\,\%$, je oznaka enaka **,
- kjer je $1\,\% < p$-vrednost $\leq 5\,\%$, je oznaka enaka *,
- kjer je $p$-vrednost $> 5\,\%$, pišemo neznačilen.

Za izračun intervala zaupanja in $p$-vrednosti smo uporabili t.i. $Z$-test, ki se uporablja pri testiranju hipotez, kjer je statistika pod ničelno hipotezo normalno porazdeljena slučajna spremenljivka, kar asimptotsko velja tudi v našem primeru. V tabeli 6.3 stolpca $2,5\,\%$ in $97,5\,\%$ predstavljata spodnji in zgornji meji $95\,\%$ intervala zaupanja. V stolpcu $\hat{\gamma}$ so predstavljeni parametri, ki se izračunajo kot $e^{\hat{\beta}_i}$, saj

Tabela 6.3: Model za škodno pogostost – ocena parametrov

| Nivo faktorja | $\hat{\gamma}$ | 2,5 % | 97,5 % | $p$-vrednost | značilnost |
|---|---|---|---|---|---|
| ***(Intercept)*** | 0,1527 | 0,1409 | 0,1655 | 0,00 % | *** |
| *veh_body: HBACK* | 0,9415 | 0,8749 | 1,0132 | 10,72 % | neznačilen |
| *veh_body: STNWG* | 1,0386 | 0,9636 | 1,1193 | 32,20 % | neznačilen |
| *veh_body: UTE* | 0,8268 | 0,7270 | 0,9403 | 0,38 % | ** |
| *veh_body: TRUCK* | 0,9705 | 0,8111 | 1,1612 | 74,33 % | neznačilen |
| *veh_body: HDTOP* | 1,1111 | 0,9320 | 1,3246 | 24,01 % | neznačilen |
| *veh_body: PANVN* | 1,0686 | 0,8378 | 1,3628 | 59,32 % | neznačilen |
| *veh_body: COUPE* | 1,5378 | 1,2185 | 1,9408 | 0,03 % | *** |
| *veh_body: MIBUS* | 0,9608 | 0,7132 | 1,2942 | 79,23 % | neznačilen |
| *veh_body: MCARA* | 1,7859 | 1,0737 | 2,9704 | 2,55 % | * |
| *veh_body: CONVT* | 0,5531 | 0,1782 | 1,7170 | 30,55 % | neznačilen |
| *veh_body: BUS* | 2,5324 | 1,3592 | 4,7182 | 0,34 % | ** |
| *veh_body: RDSTR* | 1,5079 | 0,4855 | 4,6833 | 47,75 % | neznačilen |
| *veh_age: 4* | 0,9241 | 0,8565 | 0,9971 | 4,18 % | * |
| *veh_age: 2* | 1,1358 | 1,0543 | 1,2236 | 0,08 % | *** |
| *veh_age: 1* | 1,0887 | 1,0005 | 1,1846 | 4,87 % | * |
| *agecat: 3* | 1,0301 | 0,9503 | 1,1167 | 47,09% | neznačilen |
| *agecat: 2* | 1,0935 | 1,0049 | 1,1898 | 3,81 % | * |
| *agecat: 5* | 0,8035 | 0,7301 | 0,8843 | 0,00 % | *** |
| *agecat: 6* | 0,8132 | 0,7249 | 0,9122 | 0,04 % | *** |
| *agecat: 1* | 1,2960 | 1,1688 | 1,4370 | 0,00 % | *** |

je eksponentna funkcija inverz logaritemske povezovalne funkcije, $\hat{\beta}_i$ pa so ocenjeni parametri. Rezultat, ki ga vidimo v vrstici »(Intercept)«, se nanaša na nivoje, ki so v izhodiščnem razredu (interceptu). V našem modelu so to tisti nivoji, ki imajo največjo izpostavljenost (*exposure*), saj tako zagotovimo večjo statistično značilnost (pri *veh_body* je to nivo SEDAN, pri *veh_age* nivo 3 in pri *agecat* nivo 4). Torej imajo zavarovanci v izhodiščnem razredu škodno pogostost $15,27\,\%$, interval zaupanja pa je enak $(14,09\,\%,\ 16,55\,\%)$. Za ostale nivoje, ki niso v izhodiščnem razredu, so podana zgolj relativna razmerja. V izhodiščnem razredu so torej tisti nivoji, ki niso navedeni v tabeli 6.3. Za zavarovanca, ki bi se od zavarovanca, ki je v izhodiščnem razredu, razlikoval v faktorju starosti zavarovanca (*agecat*) in bi namesto nivoja 4 imel npr. nivo 5, bi škodno pogostost izračunali kot $15,27\,\% \times 0,8035 = 12,27\,\%$. Interval zaupanja za njegovo škodno pogostost bi bil enak $(11,16\,\%,\ 13,49\,\%)$. Za izračun intervala zaupanja smo predpostavili, da so ocenjeni parametri $\hat{\beta}_i$ (asimptotično) normalno porazdeljene slučajne spremenljivke. V tem primeru velja, da je vsota dveh parametrov ponovno normalno porazdeljena slučajna spremenljivka, kjer je matematično upanje enako vsoti posameznih matematičnih upanj, varianca pa je enaka vsoti posameznih varianc povečana za dvakratnik kovariance. Iz tega lahko nato izračunamo standardni odklon ter interval zaupanja.

Vidimo lahko, da so nekateri nivoji statistično neznačilni ($p$-vrednost večja od $5\,\%$),

kar pomeni, da razlika med tem nivojem in nivojem v izhodiščnem razredu ni statistično značilna. V tem primeru bi lahko nivo združili z nivojem v izhodiščnem razredu, s tem pa bi mogoče zagotovili tudi večjo statistično značilnost modela.

## 6.2 Modeliranje povprečne škode

Zgornji postopek ponovimo še za povprečno škodo. Za modeliranje povprečne škode smo izbrali gama porazdelitev z logaritemsko povezovalo funkcijo. Pri povprečni škodi nas zanimajo zgolj police, ki so imele škodo, za utež $w$ pa vzamemo število škodnih dogodkov (*numclaims*). Ponovno si izberemo, da je stopnja tveganja $\alpha = 5\,\%$.

Najprej se odločimo, katere faktorje bomo vključili v model. Če vključimo vseh 5 faktorjev, dobimo naslednje rezultate

Tabela 6.4: Model za povprečno škodo – vključenih vseh 5 faktorjev

| **Faktor** | $\hat{\chi}^2$ | **Df** | ***p*-vrednost** |
|---|---|---|---|
| ***veh_body*** | 15,73 | 12 | 20,39 % |
| ***veh_age*** | 4,22 | 3 | 23,91 % |
| ***gender*** | 10,44 | 1 | 0,12 % |
| ***area*** | 15,43 | 5 | 0,87 % |
| ***agecat*** | 15,44 | 5 | 0,86 % |

Vidimo lahko, da sta tu dva faktorja statistično neznačilna (*veh_body*, *veh_age*). S postopnim izločanjem faktorjev ugotovimo, da moramo iz modela izločiti oba neznačilna faktorja, da dobimo model, v katerem so vsi faktorji statistično značilni. Rezultati so predstavljeni v tabeli 6.5.

Tabela 6.5: Model za povprečno škodo – vključeni statistično značilni faktorji

| **Faktor** | $\hat{\chi}^2$ | **Df** | ***p*-vrednost** |
|---|---|---|---|
| ***gender*** | 10,90 | 1 | 0,10 % |
| ***area*** | 14,56 | 5 | 1,24 % |
| ***agecat*** | 17,26 | 5 | 0,40 % |

Vidimo, da imamo lahko v modelu škodne pogostosti druge faktorje kot pa v modelu povprečne škode.

Ko imamo izbrane faktorje, ki jih bomo vključili v model, ocenimo parametre. Rezultati so zbrani v tabeli 6.6.

V izhodiščnem razredu imajo zavarovanci povprečno škodo enako 1.723, 24 \$, interval zaupanja pa je enak (1.508, 65 \$, 1.968, 34 \$). V izhodiščnem razredu so nivoji: nivo F pri *gender*, nivo C pri *area* in nivo 3 pri *agecat*. To so tisti nivoji, ki imajo največjo izpostavljenost, kar pri modeliranju povprečne škode pomeni, da imajo največje število škodnih dogodkov (*numclaims*). Za nivoje, ki niso v izhodiščnem

Tabela 6.6: Model za povprečno škodo – ocena parametrov

| **Nivo faktorja** | $\hat{\gamma}$ | **2,5 %** | **97,5 %** | ***p*-vrednost** | **značilnost** |
|---|---|---|---|---|---|
| ***(Intercept)*** | 1.723,24 | 1.508,65 | 1.968,34 | 0,00 % | *** |
| ***gender: M*** | 1,1863 | 1,0721 | 1,3127 | 0,10 % | *** |
| ***area: A*** | 0,9051 | 0,7896 | 1,0376 | 15,26 % | neznačilen |
| ***area: B*** | 0,9086 | 0,7879 | 1,0477 | 18,73 % | neznačilen |
| ***area: D*** | 0,9166 | 0,7669 | 1,0955 | 33,85 % | neznačilen |
| ***area: E*** | 1,0701 | 0,8804 | 1,3005 | 49,63 % | neznačilen |
| ***area: F*** | 1,3067 | 1,0469 | 1,6310 | 1,81 % | * |
| ***agecat: 4*** | 1,0035 | 0,8687 | 1,1593 | 96,21 % | neznačilen |
| ***agecat: 2*** | 1,0950 | 0,9419 | 1,2729 | 23,76 % | neznačilen |
| ***agecat: 5*** | 0,9047 | 0,7620 | 1,0741 | 25,29 % | neznačilen |
| ***agecat: 1*** | 1,3321 | 1,1083 | 1,6010 | 0,23 % | ** |
| ***agecat: 6*** | 0,9581 | 0,7801 | 1,1767 | 68,29 % | neznačilen |

razredu, so ponovno podana le relativna razmerja glede na nivoje v izhodiščnem razredu.

Ko imamo modelirano škodno pogostost in povprečno škodo, lahko izračunamo nevarnostno premijo. Ker se nevarnostna premija določi kot produkt med škodno pogostostjo in povprečno škodo, lahko za vsakega zavarovanca najprej ločeno določimo škodno pogostost in povprečno škodo. Za primer vzemimo zavarovanca, ki ima naslednje lastnosti: *veh_body*: SEDAN, *veh_age*: 3, *agecat*: 5, *area*: C, *gender*: M. Najprej iz modela za škodno pogostost določimo njegovo škodno pogostost. Vidimo lahko, da faktorja *area* in *gender* nista vključena v model in zato ne vplivata na izračun škodne pogostosti. Ker sta nivoja *veh_body*: SEDAN in *veh_age*: 3 v izhodiščnem razredu, moramo škodno pogostost za izhodiščni razred zgolj množiti z relativnim razmerjem za *agecat*: 5 glede na nivo *agecat*: 4, ki je v izhodiščnem razredu. Tako dobimo, da je škodna pogostost enaka $15,27\,\% \times 0,8035 = 12,27\,\%$. Sedaj izračunajmo še povprečno škodo. Vidimo lahko, da v modelu za povprečno škodo ni faktorjev *veh_body* in *veh_age*, torej ne vplivata na izračun povprečne škode. Nivo *area*: C je v izhodiščnem razredu, povprečno škodo za izhodiščni razred pa moramo sedaj množiti še z relativnima razmerjema za *agecat*: 5 in *gender*: M, saj sta v izhodiščnem razredu nivoja *agecat*: 3 in *gender*: F. Povprečna škoda za zavarovanca je enaka $1.723,24\,\$ \times 1,1863 \times 0,9047 = 1.849,46\,\$$. Nevarnostna premija za tega zavarovanca je enaka $12,27\,\% \times 1.849,46\,\$ = 226,93\,\$$.

# 7 Zaključek

GLM je v zavarovalništvu zelo uporaben model, saj lahko z njim ocenimo rizičnosti zavarovanca, s tem pa tudi ocenimo, kolikšna mora biti njegova nevarnostna premija, da bo zadostovala za pokritje bodočih škod. Zavarovalnica lahko z uporabo GLM pri oblikovanju cen poskrbi za večjo konkurenčnost, hkrati pa izboljša svoj portfelj, saj lahko zavarovancem ponudi premijo, ki bo sorazmerna z njegovo rizičnostjo. Velika prednost GLM je tudi v tem, da s testiranjem hipotez zavarovalnica oceni, kako dobro se postavljeni model prilega podatkom oziroma lahko posamezne modele med seboj primerja in tako izbere tistega, ki je najbolj primeren za trg, na katerem deluje.

Uporaba GLM pri oblikovanju cen je po svetu že zelo razširjena. Zaradi tega so se razvili tudi različni programi, ki omogočajo hitrejšo in bolj priročno uporabo GLM pri oblikovanju cen. Vendar pa je za začetek dovolj tudi uporaba programskega orodja R, kar smo videli tudi v primeru.

Za enkrat se zdi, da v Sloveniji uporaba GLM še ni tako razširjena kot v drugih bolj razvitih državah Evrope, vendar pa bosta ravno padanje povprečne premije in velika konkurenčnost zavarovalnic poskrbeli, da se bo uporaba GLM začela še bolj širiti in nadgrajevati. Ravno zaradi tega lahko pričakujemo, da se bo v prihodnjih letih z uporabo GLM zelo spremenila struktura zavarovalnih premij. Verjamemo, da bo takrat zelo pomembno, da bodo slovenske zavarovalnice dovolj hitro sledile spremembam, saj bodo le tako lahko ostale konkurenčne in solventne.

# Literatura


[1] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher in N. Thandi, *A Practitioner's Guide to Generalized Linear Models*, 3. izd., Towers Watson, London, 2007, [ogled 12. 10. 2014], dostopno na `http://www.towerswatson.com/en/Insights/IC-Types/Technical-Regulatory/2010/A-Practitioners-Guide-to-Generalized-Linear-Models`.

[2] R.E. Beard, T. Pentikäinen in E. Pesonen, *Risk Theory*, Chapman & Hall, London, 1984.

[3] M. Bijelić, *Zavarovanje in pozavarovanje*, Art agencija, Ljubljana, 1998.

[4] J. Boncelj, *Zavarovalna ekonomika*, Založba Obzorja, Maribor, 1983.

[5] B. Jørgensen, *The Theory of Exponential Dispersion Models and Analysis of Deviance*, IMPA, Rio de Janeiro, 1992, [ogled 2. 1. 2015], dostopno na `http://www.impa.br/opencms/pt/biblioteca/mono/Mon_51.pdf`.

[6] M. Lamovšek, T. Trampuž in T. Korbar, *Zavarovanja motornih vozil 2015*, Slovensko zavarovalno združenje, Ljubljana, 2015, [ogled 20. 12. 2014], dostopno na `http://www.zav-zdruzenje.si/portfolio/zavarovanja-motornih-vozil-2015/`.

[7] P. Močivnik, *Razlaga zavarovalniških izrazov*, Slovensko zavarovalno združenje, 2010, [ogled 15. 12. 2014], dostopno na `http://www.zav-zdruzenje.si/slovar-zavarovalnih-izrazov/`.

[8] *Obligacijski zakonik (uradno prečiščeno besedilo) (OZ-UPB1)*, Uradni list Republike Slovenije, št. 97/2007.

[9] E. Ohlsson in B. Johansson, *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag, Berlin, 2010.

[10] *Statistični zavarovalniški bilten 2014*, Slovensko zavarovalno združenje, Ljubljana, 2014, [ogled 20. 12. 2014], dostopno na `http://www.zav-zdruzenje.si/portfolio/statisticni-zavarovalniski-bilten-2014/`.

[11] *Učbenik za zavarovalne zastopnike in zavarovalne posrednike: 1.del: Zavarovalne, etične in pravne osnove*, Slovensko zavarovalno združenje, Ljubljana, 2013.