UNIVERZA V MARIBORU

FAKULTETA ZA NARAVOSLOVJE IN MATEMATIKO

Oddelek za matematiko in računalništvo

MAGISTRSKO DELO

Tjaša Navotnik

Maribor, 2020

UNIVERZA V MARIBORU

FAKULTETA ZA NARAVOSLOVJE IN MATEMATIKO

Oddelek za matematiko in računalništvo

Magistrsko delo

UPORABA POSPLOŠENIH LINEARNIH MODELOV PRI ANALIZI ZAVAROVALNIŠKIH PODATKOV

na študijskem programu 2. stopnje Matematika

Mentor:

izr. prof. dr. Marko Jakovac

Somentor:
Mihael Pisanec

Kandidatka:

Tjaša Navotnik

Maribor, 2020

ZAHVALA

Hvala mentorju, izr. prof. dr. Marku Jakovcu in somentorju Mihaelu Pisancu za strokovno pomoč in usmerjanje pri nastajanju magistrskega dela. Najlepša hvala tudi Kristjanu Bolčiču za vse koristne nasvete pri modeliranju.

Posebna hvala družini za spodbudne besede in finančno podporo skozi celoten študij.

Uporaba posplošenih linearnih modelov pri analizi zavarovalniških podatkov

program magistrskega dela

Posplošeni linearni modeli (GLM) predstavljajo nadgradnjo običajnega linearnega modela, s katerim ponazorimo linearni odnos med eno odvisno in več neodvisnimi spremenljivkami. Linearna regresija temelji na pogojni porazdelitvi odvisne spremenljivke glede na dane vrednosti neodvisnih spremenljivk, zato je linearni model posebej učinkovit na podatkih, ki so porazdeljeni (približno) normalno. V primeru podatkov, ki niso porazdeljeni normalno, pa pride do izraza uporaba posplošenih linearnih modelov. Uporaba teh modelov je v zavarovalništvu že zelo razširjena in se praviloma uporablja pri neživljenjskih zavarovanjih (npr. avtomobilska zavarovanja). V magistrskem delu bo predstavljena teorija posplošenih linearnih modelov in njihova uporaba pri obdelavi zavarovalniških podatkov. Prikazani bodo primeri uporabe teh modelov na različnih tipih neživljenjskih zavarovanj, prav tako pa bo napravljena analiza, kako bi te modele lahko uporabili še pri življenjskih zavarovanjih. V zaključku magistrskega dela bo narejen še izračun premij za nekatera zavarovanja na realnih podatkih izbrane zavarovalnice.

Osnovni viri:

1. A. J. Dobson, A. G. Barnett, *An Introduction to Generalized Linear Models*, 4. edition, Chapman and Hall/CRC, 2018.
2. P. de Jong, G. Z. Heller, *Generalized Linear Models for Insurance Data*, Cambridge University Press, 2008.
3. E. Ohlsson, B. Johansson, *Non-life Insurance Pricing with Generalized Linear Models*, Springer-Verlag Berlin Heidelberg, 2010.

izr. prof. dr. Marko Jakovac

Mihael Pisanec

NAVOTNIK, T. : Uporaba posplošenih linearnih modelov pri analizi zavarovalniških podatkov.
Magistrsko delo, Univerza v Mariboru, Fakulteta za naravoslovje in matematiko, Oddelek za matematiko in računalništvo, 2020.

IZVLEČEK

V magistrskem delu je predstavljena teorija posplošenih linearnih modelov (GLM) in uporaba le-teh v zavarovalništvu. Na realnih zavarovalniških podatkih izbrane zavarovalnice s pomočjo modelov GLM segmentiramo zavarovance glede na rizičnost in s tem določimo premijo za zavarovanje avtomobilske odgovornosti, ki je prilagojena posamezniku. S pomočjo logistične regresije analiziramo prekinitve življenjskih zavarovanj in zgradimo klasifikacijski model, ki razvrsti zavarovance v dve skupini; na tiste, ki so oz. niso nagnjeni k prekinitvi.

Ključne besede: posplošeni linearni modeli, avtomobilsko zavarovanje, prekinitve življenjskih zavarovanj, modeliranje v R.

Math. Subj. Class. (2010): 62J12, 62P05.

NAVOTNIK, T. : Applying generalized linear models to insurance data analysis.
Master's Thesis, University of Maribor, Faculty of Natural Sciences and Mathematics, Department of Mathematics and Computer Science, 2020.

ABSTRACT

The master's thesis presents the theory of generalized linear models (GLM) and their application in the insurance industry. Based on the real insurance data of the selected insurance company, we use the GLM models to segment insured persons according to risk, thus determining the premium for individual motor third party liability insurance. We use logistic regression to analyze the termination of life insurance policies and we build a classification model that classifies policyholders into two groups; to those who are and who are not prone to lapse.

Keywords: generalized linear models, motor insurance, termination of life insurance policies, modeling in R.

Math. Subj. Class. (2010): 62J12, 62P05.

Kazalo

Uvod

Za konkurenčnost na zavarovalniškem trgu je ključna segmentacija strank glede na njihovo rizičnost. To lahko dosežemo z uporabo posplošenih linearnih modelov (v nadaljevanju GLM), ki so v zavarovalništvu široko uporabljeni predvsem pri neživljenjskih zavarovanjih in so tema tega magistrskega dela.

Delo je razdeljeno na tri glavna poglavja. V prvem delu razložimo osnove verjetnosti in statistike ter osnove zavarovalništva, predstavimo določitev premij za neživljenjska zavarovanja, opišemo delitev življenjskih zavarovanj ter prekinitev le-teh.

V drugem delu je poudarek na teoriji posplošenih linearnih modelov. Najprej predstavimo linearni regresijski model, njegove predpostavke in podamo razloge, zakaj ni povsem primeren za zavarovalniške podatke, nato opišemo posplošene linearne modele in delo z njimi.

V zadnjem delu se posvetimo uporabi teh modelov na realnih zavarovalniških podatkih izbrane zavarovalnice. Za zavarovanje avtomobilske odgovornosti (AO) s pomočjo modeliranja škodne pogostosti in višine škode določimo premijo, ki je prilagojena posamezniku in temelji na pravičnosti. Predstavimo vse korake modeliranja, od predanalize podatkov, izbire pojasnjevalnih spremenljivk, do preverjanja ustreznosti modela. To premijo primerjamo s trenutnimi premijami zavarovalnice ter s tem pokažemo, v katerih segmentih zavarovalnica že zaračunava teoretično ustrezne premije in kje bi bile potrebne modifikacije. Uporabo GLM nato apliciramo še na področje življenjskega aktuarstva in sicer z analizo prekinitev življenjskih zavarovanj. S simulacijo odvisne spremenljivke ustvarimo enega izmed možnih scenarijev in s pomočjo logistične regresije analiziramo vpliv posameznih pojasnjevalnih spremenljivk na predčasno prekinitev sklenjenega življenjskega zavarovanja. Na koncu preverimo napovedno moč modela.

Podatki so v samem delu prikazani le delno zaradi varovanja podatkov zavarovalnice. Statistične analize so izvedene z uporabo programskega orodja RStudio, ki uporablja programski jezik R.

Poglavje 1

Verjetnost in statistika

Najprej definirajmo pojme iz verjetnosti in statistike, ki se bodo pojavljali skozi celotno magistrsko delo. Vsebina poglavja je vzeta iz literatur [6, 23].

Definicija 1.1 *Naj bo G prostor elementarnih dogodkov. Neprazna družina podmnožic $\mathcal{A}$ v G je σ-algebra, če velja*

i) $A \in \mathcal{A} \Rightarrow \overline{A} \in \mathcal{A}$,

i) za vsako družino $\{A_n | n \in \mathbb{N}\} \subseteq \mathcal{A}$, *je tudi* $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}$.

Definicija 1.2 *Borelova σ-algebra na $\mathbb{R}$ je σ-algebra $\mathcal{B}$, ki je generirana z odprtimi množicami (t.i. Borelovimi množicami).*

Opomba 1.3 *$\mathcal{B}$ je najmanjša σ-algebra na $\mathbb{R}$, ki vsebuje vse odprte množice.*

Naključna spremenljivka je količina, ki v vsakem poskusu zavzame neko realno vrednost. Katero vrednost zavzame, pa je odvisno od naključja. Zapišimo še formalno definicijo naključne oz. slučajne spremenljivke.

Definicija 1.4 *Naj bo $(G, \mathcal{A}, P)$ verjetnostni prostor. Funkcija $X : G \to \mathbb{R}$ je naključna spremenljivka, če za vsako Borelovo množico $B \in \mathcal{B}$ velja $X^{-1}(B) \in \mathcal{A}$.*

Definicija 1.5 *Naj bo X naključna spremenljivka. Porazdelitvena funkcija $F_X : \mathbb{R} \to [0, 1]$ naključne spremenljivke X je za vsak $x \in \mathbb{R}$ definirana kot $F_X(x) = P(X < x)$.*

Definicija 1.6 *X je diskretna naključna spremenljivka, če je njena zaloga vrednosti števna množica.*

Naj bo $Z_X = \{x_1, x_2, x_3, \dots\}$ zaloga vrednosti diskretne naključne spremenljivke. Splošna porazdelitev je za vsak $i \in \mathbb{N}$ podana z verjetnostno funkcijo

$$p_i = P(X = x_i)$$

in velja $\sum_{i=1}^{\infty} p_i = 1$.

Definicija 1.7 *Naj bo X naključna spremenljivka in F_X njena porazdelitvena funkcija. Pravimo, da je X porazdeljena zvezno, če obstaja nenegativna realna funkcija $p : \mathbb{R} \to \mathbb{R}$, da za vsak $x \in \mathbb{R}$ velja $F_X(x) = \int_{-\infty}^{x} p(t)dt$. Funkcija p je gostota verjetnosti.*

Opomba 1.8 *Pogosto funkcijo verjetnosti oz. gostoto verjetnosti naključne spremenljivke X označimo tudi s f_X.*

Pri naključnih spremenljivkah nas poleg porazdelitvene funkcije pogosto zanimajo številske karakteristike le-teh. Pod osnovne številske karakteristike naključnih spremenljivk spadajo matematično upanje, varianca, kovarianca, pogojno matematično upanje in momenti. Slednji so posplošitev matematičnega upanja in variance. Definirali jih bomo v poglavju 3.4.1.

Definicija 1.9 *Naj bo X naključna spremenljivka s porazdelitveno funkcijo F_X. Če obstaja $\int_{\mathbb{R}} |x|dF$, potem definiramo $E(X) = \int_{\mathbb{R}} |x|dF$. Vrednost $E(X)$ imenujemo matematično upanje (pričakovana vrednost, povprečje) in predstavlja pričakovano vrednost v zalogi naključne spremenljivke X. Natančneje:*

1. *Naj bo X diskretna naključna spremenljivka z verjetnostno funkcijo $p_i = P(X = x_i)$; $i \in \mathbb{N}$. Če konvergira vrsta $\sum_{i=1}^{\infty} |x_i|p_i < \infty$, potem je $E(X) = \sum_{i=1}^{\infty} x_i p_i$.*
2. *Naj bo X zvezna naključna spremenljivka z gostoto verjetnosti p. Če obstaja integral $\int_{\mathbb{R}} |x|p(x)dx$, potem je $E(X) = \int_{\mathbb{R}} xp(x)dx$.*

Opomba 1.10 *Integral, uporabljen v definiciji 1.9, je Rieman-Stieltjesov integral.*

Naj bo $f : [a, b] \to \mathbb{R}$ omejena funkcija in $F : [a, b] \to \mathbb{R}$ naraščajoča funkcija. Naj bo $D = \{x_0, x_1, x_2, \dots, x_n\}$, $a = x_0 < x_1 < x_2 < \cdots < x_n = b$, delitev intervala $[a, b]$. Označimo $\Delta = \{\max |x_k - x_{k-1}| \mid k = 1, \dots, n\}$. Na vsakem podintervalu $[x_{k-1}, x_k]$ izberimo

poljubno točko s_k. Riemann–Stieltjesov integral funkcije f glede na funkcijo F je definiran kot

$$\int_a^b f dF = \lim_{\Delta \to 0} \sum_{k=1}^{n} f(s_k) \left(F(x_k) - F(x_{k-1})\right).$$

V posebnem primeru, ko je $F(x) = x$, dobimo Riemmanov integral

$$\int_a^b f dF = \lim_{\Delta \to 0} \sum_{k=1}^{n} f(s_k)(x_k - x_{k-1}).$$

Definicija 1.11 *Za vsako naključno spremenljivko X, ki ima matematično upanje $E(X)$, je njena disperzija ali varianca definirana kot*

$$D(X) = \mathrm{Var}(X) = E((X - E(X))^2)$$

in predstavlja mero za razpršenost porazdelitve okoli pričakovane vrednosti. Število $\sigma(X) = \sqrt{\mathrm{Var}(X)}$ se imenuje standardni odklon.

Varianca je lahko končna ali neskončna, je pa vedno $\mathrm{Var}(X) \geq 0$. Lahko jo zapišemo tudi v obliki $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

Definicija 1.12 *Kovarianca med naključnima spremenljivkama X in Y je definirana kot*

$$\mathrm{Kov}(X, Y) = E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y).$$

Če je $\mathrm{Kov}(X, Y) = 0$, pravimo, da sta X in Y nepovezani (nekolerirani) naključni spremenljivki.

Definicija 1.13 *Naj bo X naključna spremenljivka na $(G, \mathcal{A}, P)$ in $A \in \mathcal{A}$; $P(A) > 0$. Pogojna naključna spremenljivka $X|A$ je naključna spremenljivka s porazdelitveno funkcijo*

$$F_{X|A} = P((X < x)|A) = \frac{P((X < x)A)}{P(A)}.$$

Definicija 1.14 *Pogojno matematično upanje je matematično upanje pogojne naključne spremenljivke $X|A$: $E(X|A) = \int_\infty x dF_{X|A}$.*

Poglejmo si nekaj pomembnih porazdelitev, s katerimi se bomo srečevali v nadaljevanju.

Pomembni diskretni porazdelitvi:

1. Binomska porazdelitev $B(n,p)$.
   Naj ima dogodek A v nekem poskusu verjetnost $p = P(A)$; $p \in (0,1)$. Potem je $P(\bar{A}) = 1-p = q$. Dani poskus neodvisno ponovimo n-krat ($n \in \mathbb{N}$). Naključna spremenljivka X naj predstavlja frekvenco dogodka A. Zaloga naključne spremenljivke je $Z_X = \{0,1,\ldots,n\}$; verjetnostna funkcija je $p_k = P(X=k) = \binom{n}{k} p^k q^{n-k}$; $k \in \mathbb{N}_0$. Matematično upanje je $E(X) = np$, varianca pa je enaka $\mathrm{Var}(X) = np(1-p)$.

2. Poissonova porazdelitev $Poiss(\lambda)$.
   Naj bo $\lambda > 0$. Zaloga vrednosti naključne spremenljivke X, ki je porazdeljena po zakonu $Poiss(\lambda)$, je množica $\mathbb{N}_0$, verjetnostna funkcija pa $p_k = P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda}$ za vsak $k \in \mathbb{N}_0$.

   Za to porazdelitev velja $\mathrm{Var}(X) = E(X) = \lambda$. Poissonova porazdelitev se najpogosteje uporablja za štetje dogodkov, ki se zgodijo v določenem času, služi pa tudi kot aproksimacija za binomsko porazdelitev.

Pomembne zvezne porazdelitve:

1. Normalna (Gaussova) porazdelitev $N(\mu, \sigma^2)$.
   Pravimo, da je X porazdeljena po normalnem zakonu $N(\mu,\sigma^2)$, kjer je μ matematično upanje in σ standardni odklon, če ima njena gostota verjetnosti predpis

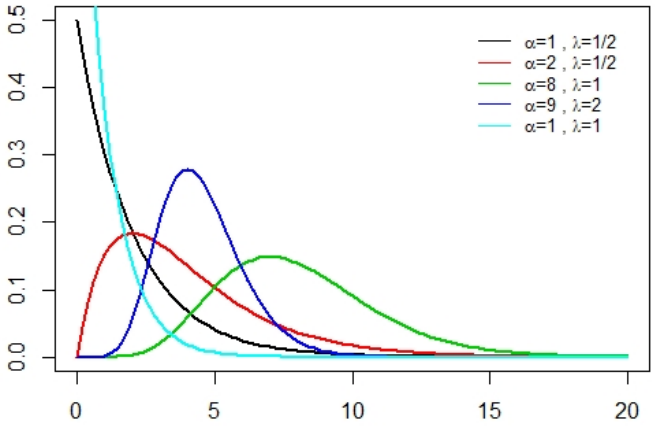
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2};\ x \in \mathbb{R}.$$

2. Gama porazdelitev $\Gamma(\alpha,\lambda)$.
   Naj bosta $\alpha > 0$ (parameter oblike) in $\lambda > 0$ (parameter stopnje). Naključna spremenljivka je porazdeljena po gama porazdelitvi, $X \sim \Gamma(\alpha,\lambda)$, če ima njena gostota verjetnosti predpis

$$p(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}; & x > 0, \\ 0; & x \le 0, \end{cases}$$

   kjer je $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$ predpis Eulerjeve funkcije gama. Matematično upanje je enako $\frac{\alpha}{\lambda}$, varianca pa je enaka $\frac{\alpha}{\lambda^2}$. Uporablja se za pozitivne zvezne spremenljivke z asimetrično porazdelitvijo (slika 1.1).

   Poseben primer te porazdelitve je porazdelitev $\chi^2(n) = \Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$, kjer je $n \in \mathbb{N}$ in pomeni število prostostnih stopenj.



Slika 1.1: Funkcija gostote verjetnosti za porazdelitev gama za različne parametre.

3. Inverzna Gaussova porazdelitev $IG(\mu, \lambda)$.
   Naključna spremenljivka X je porazdeljena po inverzni Gaussovi porazdelitvi $IG(\mu, \lambda)$, kjer je $\mu > 0$ matematično upanje in $\lambda > 0$ parameter oblike, če ima njena gostota verjetnosti predpis

$$p(x) = \frac{\lambda}{\sqrt{2\pi x^3}} e^{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}}; \; x > 0.$$

   Velja $E(X) = \mu$ in $\mathrm{Var}(X) = \frac{\mu^3}{\lambda}$. Porazdelitev je podobna gama porazdelitvi, vendar ima v primerjavi z gama porazdelitvijo ostrejši vrh in širši rep.

Poglavje 2

Osnove zavarovalništva

Osnovni cilj zavarovalništva je zaščita pred negativnimi finančnimi posledicami. Zavarovalnica pripravi za zavarovanca [1] pogodbo (zavarovalno polico), s katero bo pokrila morebitne stroške v primeru nastanka škode. Vrednost, ki jo zavarovanec terja od zavarovalnice oz. jo zavarovalnica plača v primeru škode, imenujemo zavarovalnina oz. višina zahtevka. Premija je vrednost, ki jo zavarovanec plačuje v zameno za zavarovanje. Zavarovalnica mora določiti takšno višino premije, da bo ugodna za zavarovanca in hkrati finančno vzdržna za zavarovalnico. Zavarovalna polica je torej dogovor med zavarovalnico in zavarovancem, s katerim zavarovalnica prevzame finančno odgovornost za morebitno nastalo škodo predmeta zavarovanja oz. zavarovalnega dogodka v nekem določenem času, zavarovalni dobi.

Zaradi zakona velikih števil je izguba zavarovalnice, ki je vsota velikega števila relativno majhnih neodvisnih izgub, veliko bolj predvidljiva, kot izguba posameznika. Ta namreč zagotavlja, da je praktična vrednost blizu teoretične oz. zagotavlja stabilnost podatkov na dolgi rok (ob ustreznih predpostavkah). Povedano zapišimo še formalno.

Izrek 2.1 (Krepki zakon velikih števil) *Naj bo $(X_n)_{\mathbb{N}}$ zaporedje neodvisnih in identično porazdeljenih naključnih spremenljivk z matematičnim upanjem μ. Za vsak $i \in \mathbb{N}$ vpeljemo novo naključno spremenljivko $\overline{X_i} = \frac{X_1+X_2+\cdots+X_i}{i}$. Tedaj velja $P\left(\lim_{i\to\infty} \overline{X_i} = \mu\right) = 1$.*

To nas pripelje do splošno sprejetega principa, da mora premija temeljiti na pričakovani (povprečni) izgubi, ki se prenese z zavarovanca na zavarovalnico. Nato se k tej neto premiji μ doda znesek, s katerimi se pokrijejo začetna provizija (sklenitveni stroški), upravni in

[1] Zavarovanec je oseba, ki je zavarovana glede na zavarovalni interes. Zavarovalec je oseba, ki sklene zavarovalno pogodbo. Zavarovalec in zavarovanec sta pogosto ista oseba, zato med njima ne bomo razlikovali.

drugi stroški (tekoči stroški za plače, takse, ipd.) ter stroški poravnave zahtevkov. S tem dobimo tako imenovano bruto oz. kosmato premijo, ki ima naslednjo obliko:

$$H = (1 + a)\mu + b;\ a, b > 0, \tag{2.1}$$

kjer so $a\mu$ stroški vezani na premijo (varnostni dodatek), b pa so stroški neodvisni od premije.

Pričakovana izguba variira med policami: število nesreč ni enako za vse zavarovance in ko pride do zahtevka, se pričakovana škoda razlikuje med njimi. Torej je smiselno, da je zavarovanje proti požaru za razkošno vilo višje kot za manjšo hišo, da vozniki, ki povzročijo več prometnih nesreč plačajo več za avtomobilsko zavarovanje oz. po drugi strani varni vozniki dobijo ugodnost v obliki zmanjšane premije. Prav tako bi namestitev protipožarnega alarma, ki zmanjšuje pogostost zahtevka, lahko omogočila popust pri premiji za zavarovanje hiš. S tem ko zavarovalnice upoštevajo te dejavnike in določijo nižjo premijo za ugodnega zavarovanca, lahko bolj konkurirajo na trgu in pridobijo več profitabilnih zavarovancev. Hkrati izgubijo nekatere rizične zavarovance, saj morajo ti plačevati višji znesek za zavarovanje in si s tem izboljšajo kvaliteto portfelja. Zavarovalnicam je torej v interesu, da zaračunajo pošteno premijo, to pomeni, da zavarovanec, kolikor je to mogoče, plača premijo, ki ustreza pričakovanim izgubam, ki jih je prenesel na zavarovalnico [14].

Takšno oblikovanje cen lahko določimo s pomočjo generaliziranimi linearnimi modeli, ki se jim bomo posvetili v naslednjem poglavju. Najprej pa razložimo osnovne pojme iz zavarovalništva, s katerimi se bomo srečevali skozi to delo.

2.1 Določitev premije za neživljenjska zavarovanja

Pri določanju neto premije za premoženjska zavarovanja pogosto analiziramo škodno pogostost in povprečno višino zahtevka (škode).

Škodna pogostost predstavlja povprečno število zahtevkov na enoto izpostavljenosti (npr. na časovno enoto, ponavadi eno leto) [14] in se izračuna kot

$$\text{škodna pogostost} = \frac{\text{število zahtevkov}}{\text{število enot izpostavljenosti}}. \tag{2.2}$$

Povprečna škoda predstavlja povprečno vrednost zahtevka in je enaka celotnemu znesku, ki ga plača zavarovalnica, deljenim s številom prijavljenih zahtevkov:

$$\text{povprečna škoda} = \frac{\text{višina izplačanih škod}}{\text{število zahtevkov}}. \tag{2.3}$$

Neto premija je produkt škodne pogostosti in povprečne škode:

$$\text{neto premija} = \text{škodna pogostost} \cdot \text{povprečna škoda}. \tag{2.4}$$

2.2 Življenjska zavarovanja

Življenjska zavarovanja se navezujejo na zavarovanje oseb. Pri teh zavarovanjih je pomembno trajanje zavarovanja ter višina izplačila, pravimo ji zavarovalna vsota, ki se izplača ob zavarovalnem dogodku. Pri različnih zavarovalnicah lahko sklenemo različne oblike življenjskega zavarovanja. Poglejmo si nekatere pogoste oblike [20].

1. Življenjsko zavarovanje za primer smrti.
   Pri tem zavarovanju pride do izplačila dogovorjene zavarovalne vsote, če zavarovana oseba umre v času trajanja zavarovanja. Poznamo več oblik tega zavarovanja, med drugimi doživljenjsko zavarovanje, kjer je izplačilo zajamčeno (zavarovalna vsota se izplača ob smrti zavarovanca). Mogoče je skleniti zavarovanje za primer smrti, ki je časovno omejeno. Takšno zavarovanje imenujemo tudi riziko življenjsko zavarovanje. Do izplačila pride, če zavarovana oseba umre v vnaprej določenem času trajanja zavarovanja. Če le-ta umre po izteku tega obdobja, ne pride do izplačila.

2. Življenjsko zavarovanje za primer doživetja.
   Zavarovani osebi se izplača dogovorjena zavarovalna vsota, če preživi dogovorjeno dobo. Če le-ta umre med trajanjem zavarovanja, zavarovalnici vse obveznosti prenehajo.

3. Mešano življenjsko zavarovanje.
   Mešano življenjsko zavarovanje je življenjsko zavarovanje za primer smrti in doživetja. Če zavarovana oseba umre v času trajanja zavarovanja, se upravičencu [2] izplača dogovorjen znesek. Sicer se ob koncu zavarovalne dobe izplača zavarovalna vsota za primer doživetja.

4. Naložbeno življenjsko zavarovanje.
   To zavarovanje je prav tako zavarovanje za primer smrti in doživetja. Za razliko od mešanega življenjskega zavarovanja, kjer je višina izplačila ob doživetju zajamčena, pri tem zavarovanju višina zavarovalne vsote za doživetje ni natančno določena. Varčevalni del se namreč vlaga v sklade in se spreminja z indeksom, na katerega je zavarovanje vezano [25].

[2] Upravičenec je oseba, ki ji v skladu zavarovalne pogodbe pripada zavarovalna vsota. Pri zavarovanju za primer smrti je upravičenec zakoniti dedič, razen če je navedeno drugače.

2.2.1 Prekinitve življenjskih zavarovanj

Življenjska zavarovanja so običajno sklenjena na dolgi rok. V kolikor zavarovancu finančno stanje več ne dopušča plačevanja premij, ali jih zaradi drugega razloga ne želi več plačevati, lahko predčasno prekine zavarovalno pogodbo. Kdaj in na kakšen način lahko prekine zavarovanje je odvisno od posamezne zavarovalnice in od zavarovalnih pogojev, ki so navedeni v zavarovalni pogodbi ob sklenitvi zavarovanja.

Če ima zavarovanec sklenjeno življenjsko zavarovanje za primer smrti (riziko življenjsko zavarovanje) in predčasno prekine zavarovalno pogodbo (kar lahko stori kadarkoli), se plačane premije ne vrnejo.

V primeru, da ima sklenjeno življenjsko zavarovanje z varčevanjem (naložbeno življenjsko zavarovanje) in prekine zavarovanje po preteku dveh oz. treh letih zavarovanja (odvisno od zavarovalne dobe), lahko dobi izplačan del sredstev. V kolikor premija ni bila plačana za najmanj dve oz. tri leta in zavarovanec stornira zavarovanje, se do tedaj plačane premije ne vrnejo.

Po preteku dveh let oz. treh let zavarovanja (za katerega so bile plačane premije), zavarovanje lahko ostane v veljavi brez da bi zavarovanec plačeval nadaljnje premije, zavarovalna vsota pa se ustrezno zmanjša glede na število vplačanih premij in preostale dobe trajanja zavarovanja. V tem primeru pravimo, da pride do kapitalizacije [13].

Po poteku dveh oz. treh letih zavarovanja lahko zavarovanec, v kolikor je poravnal vse obveznosti do zavarovalnice, zahteva izplačilo odkupne vrednosti, predujma v določeni višini (običajno do 80 % višine odkupne vrednosti) ali poda prošnjo za mirovanje plačevanja premije. Mirovanje lahko neprekinjeno traja največ eno leto, predujem pa je potrebno vrniti z obrestmi [24].

V primeru odkupa (zavarovanec mora izpolnjevat pogoje, ki so navedeni v zavarovalnih pogojih) se zavarovanje prekine, zavarovancu pa se izplača vrednost zavarovalne police (privarčevana sredstva oziroma matematična rezervacija, znižana za stroške zavarovalnice)[13].

Prekinitev življenjskega zavarovanja je lahko posledica nezmožnosti plačevanja premije zaradi izgube službe, znižanja plače, kreditne obremenitve in podobno. Eden izmed razlogov za prekinitev je tudi nepremišljeno sklenjeno zavarovanje oz. neustrezen zavarovalni produkt, [3] npr. osebe ne potrebujejo življenjskega zavarovanja ali pa so sklenili zavarovanje za zanj neustrezno zavarovalno vsoto ali mesečno premijo. V primeru naložbenega življenjskega zavarovanja je izplačan znesek odvisen od poslovanja skladov. V kolikor investicijski skladi

[3] Življenjsko zavarovanje za primer smrti se priporoča osebam z družino, ki predstavljajo glavni vir prihodkov in želijo poskrbeti za finančno preskrbljenost ostalih članov družine po zavarovančevi smrti [19].

slabo poslujejo, se zmanjšajo tudi prihranki zavarovanca in posledično več ne želi imeti takšnega zavarovanja [12].

Prekinitve zavarovalnih polic življenjskega zavarovanja lahko ima za zavarovalnico dolgoročne posledice. Visoka stopnja prekinitev namreč negativno vpliva na ugled zavarovalnice, kar lahko povzroči še več prekinitev oz. manj novo sklenjenih zavarovalnih polic. Da bi zmanjšali tveganje prekinitev in finančnih izgub, ki bi lahko nastali v prihodnosti, poskušajo zavarovalnice oblikovati učinkovit klasifikacijski model, ki zavarovance razdeli v dve skupini: na tiste, ki so oz. niso nagnjeni k prekinitvi [9].

Poglavje 3

Posplošeni linearni modeli

3.1 Linearni modeli

Linearni (regresijski) modeli opisujejo povezavo med odvisno spremenljivko [1] y in k neodvisnimi oz. pojasnjevalnimi spremenljivkami [2] $x_1, x_2, \ldots, x_k$.

Linearni populacijski regresijski model zapišemo v obliki

$$E(y|x_{1i}, x_{2i}, \ldots, x_{ki}) = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

oz.

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i,$$

kjer je

- y_i vrednost odvisne spremenljivke pri i-ti opazovani enoti,
- x_{ji} vrednost pojasnjevalne spremenljivke x_j za opazovano enoto i, [3]
- β_j (parcialni) regresijski koeficient j-te pojasnjevalne spremenljivke,
- u_i slučajni odklon (napaka) pri i-ti opazovani enoti,

[1] Drugi izrazi za odvisno spremenljivko, ki se pojavljajo v literaturi, so pojasnjena spremenljivka, regresand, prediktand, output spremenljivka, endogena spremenljivka ter odzivna spremenljivka.

[2] Za neodvisno spremenljivko se pogosto uporablja izraz pojasnjevalna spremenljivka, zasledimo pa tudi naslednja poimenovanja: nepojasnjena spremenljivka, regresor, prediktor, input spremenljivka, eksogena spremenljivka ter kontrolna spremenljivka.

[3] Večinoma se uporablja regresijski model s konstantnim členom. To pomeni, da prva pojasnjevalna spremenljivka ni prava pojasnjevalna spremenljivka, ampak predstavlja regresijsko konstanto, njene vrednosti so pri vseh opazovanih enotah enake 1.

- n število vseh opazovanih enot in
- k število pojasnjevalnih spremenljivk.

Linearni model lahko zapišemo tudi v matrični obliki in sicer kot $Y = X\beta + u$, kjer je X matrika vrednosti (zveznih) pojasnjevalnih spremenljivk, β vektor regresijskih koeficientov in Y vektor odvisnih spremenljivk. Zapis linearnega regresijskega modela vzorčnih podatkov v matrični obliki je $y = Xb + e$, kjer je e vektor ostankov regresijskega modela in b je vektor ocenjenih vrednosti regresijskih koeficientov $\beta_1, \beta_2, \ldots, \beta_k$ [15].

3.1.1 Razlaga regresijskih koeficientov

Pri zveznih pojasnjevalnih spremenljivkah j-ti parcialni regresijski koeficient predstavlja povprečno spremembo pričakovane vrednosti odvisne spremenljivke, če se vrednost j-te pojasnjevalne spremenljivke spremeni za eno enoto, vrednosti ostalih pojasnjevalnih spremenljivk pa ostanejo nespremenjene [15].

Za kategorične pojasnjevalne spremenljivke uporabimo umetne oz. slamnate spremenljivke (angl. *dummy*). Temeljni princip je naslednji: če ima opazovana enota proučevano značilnost, potem je vrednost slamnate spremenljivke 1, v nasprotnem primeru pa 0. Če kategorična spremenljivka zavzame m različnih kategorij, potem za to spremenljivko izberemo $(m-1)$ slamnatih spremenljivk. [4] Npr. pri spolu uporabimo eno slamnato spremenljivko (dve kategoriji: moški ali ženska), pri čemer eno kategorijo izberemo kot bazno. V teoriji je bazna kategorija lahko poljubno izbrana (v primeru, da nima maloštevilno opazovanih enot), vendar je izbira kategorije, ki ima največ opazovanih enot pogosta praksa [2]. Tako dobimo matriko, ki je sestavljena iz 0 in 1 in jo imenujemo indikatorska matrika. Regresijski koeficient v tem primeru predstavlja razliko, ki jo posamezna lastnost povzroči na odvisno spremenljivko v primerjavi z bazno kategorijo.

3.1.2 Predpostavke

Linearni modeli so ocenjeni z metodo najmanjših kvadratov, ki minimizira vsoto kvadratov ostankov regresijskega modela. Da lahko uporabimo to metodo (kot cenilko regresijskih koeficientov linearnega modela) in trdimo, da je najboljša [5] med vsemi cenilkami, morajo biti izpolnjene naslednje predpostavke [15]:

[4] V primeru, da obravnavamo model z regresijsko konstanto.

[5] Metoda najmanjših kvadratov je ob zgornjih predpostavkah NeNaLiCe regresijskih koeficientov, kar pomeni, da je nepristranska, najboljša (z najmanjšo možno variaco) linearna cenilka regresijskih koeficentov linearnega populacijskega regresijskega modela.

1. Regresijski model je linearen v parametrih in je pravilno specificiran.

2. Za vsako vrednost pojasnjevalne spremenljivke $(x_{1i}, \ldots, x_{ki})$ velja, da je pričakovana vrednost slučajne spremenljivke u enaka 0; $E(u_i) = 0$. To pomeni, da je skupni učinek vseh dejavnikov, ki niso zajeti v vrednostih pojasnjevalnih spremenljivk, na povprečno vrednost odvisne spremenljivke enak 0.

3. Ničelna kovarianca med vrednostmi spremenljivke u, oz. ne obstaja avtokorelacija. Torej velja $\mathrm{Kov}(u_i, u_j) = 0, i \neq j$.

4. Varianca spremenljivke u je pozitivna konstanta in je enaka σ^2. Opravka imamo torej z enako razpršenostjo vrednosti u ne glede na to, kolikšna je vrednost pojasnjevalne spremenljivke (prisotna je homoskedastičnost).

5. Velja ničelna kovarianca med pojasnjevalno ter slučajno spremenljivko in velja

$$\mathrm{Kov}\,(x_{1i}, u_i) = \mathrm{Kov}(x_{2i}, u_i) = \cdots = \mathrm{Kov}(x_{ki}, u_i) = 0.$$

6. Med pojasnjevalnimi spremenljivkami ne obstaja popolna linearna odvisnost, torej med njimi ne obstaja natančna linearna odvisnost oblike

$$\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_j x_j + \cdots + \lambda_k x_k = 0,$$

   kjer so λ_j konstante, ne vse enake 0.

7. Slučajna spremenljivka u je normalno porazdeljena $(u_i \sim N\left(0, \sigma_u^2\right))$.

8. Število opazovanih enot mora biti večje kot število spremenljivk $(n > k)$.

9. Vrednosti x_i niso enake.

Če zgornje predpostavke niso izpolnjene, so rezultati opravljene regresijske analize napačni in posledično sprejmemo zavajajoče zaključke. Z metodo najmanjših kvradratov dobimo cenilko za regresijski koeficient, ki se izračuna kot $b = (X^T X)^{-1} X^T y$ [6]. Izpeljavo te cenilke bomo na tem mestu izpustili. [7]

[6] Oznaka T predstavlja transponiranje matrike in v nadaljevanju naloge tudi transponiranje vektorja.

[7] Metoda najmanjših kvadratov je podrobneje opisana v knjigi Osnove ekonometrije, stran 53 [15].

3.1.3 Slabosti uporabe linearnega modela na zavarovalniških podatkih

Ena izmed predpostavk linearnega modela je torej, da je slučajna spremenljivka u normalno porazdeljena ($u_i \sim N\left(0, \sigma_u^2\right)$). Posledično mora biti tudi odvisna spremenljivka Y normalno porazdeljena ($Y_i \sim N\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}, \sigma_u^2\right)$). V praksi se pri zavarovalniških podatkih pogosto pojavljajo druge porazdelitvene funkcije, npr. število škodnih zahtevkov je porazdeljeno diskretno in izdatki za škodni zahtevek so nenegativni in pogosto imajo asimetrično porazdelitev v desno stran. Druga težava linearnih modelov je ta, da je matematično upanje odvisne spremenljivke linearna funkcija pojasnjevalnih spremenljivk. Pri določanju cen premije pa so primernejši multiplikativni modeli za matematično upanje. Prav tako linearni modeli zahtevajo konstantno varianco odvisne spremenljivke, kar se v praksi izkaže, da velikokrat temu ni zadoščeno (varianca je pogosto odvisna od matematičnega upanja) [14]. Več o tem, kako odpravimo te omejitve linearnega modela, sledi v nadaljevanju.

3.2 Osnovne predpostavke modela

Predstavili bomo nekaj osnovnih predpostavk modela, ki bodo osnova za grajanje statističnega modela [14].

Predpostavka 1 (Neodvisnost zavarovalnih polic): Denimo, da imamo n različnih zavarovalnih polic. Naj Y_i označuje odvisno spremenljivko za zavarovalno polico i. Potem so $Y_i, \ldots, Y_n$ paroma neodvisne.
Predpostavka 2 (Neodvisnost časa): Denimo, da imamo n disjunktnih časovnih intervalov. Naj Y_i označuje odvisno spremenljivko v časovnem intervalu i. Potem so $Y_1, \ldots, Y_n$ paroma neodvisne.
Predpostavka 3 (Homogenost): Naj bosta poljubni zavarovalni polici v istem premijskem razredu (angl. *tariff cell*), ki imata enako prijavljeno škodo. Naj Y_i označuje odvisno spremenljivko za zavarovalno polico i. Potem imata Y_1 in Y_2 enako porazdelitveno funkcijo.

Opomba 3.1 *Odvisna spremenljivka je lahko npr. število škod (število prijavljenih zahtevkov) ali povprečna škoda.*

3.3 Osnove posplošenih linearnih modelov

Posplošene linearne modele (angl. *Generalized Linear Models - GLM*) sta leta 1972 predstavila Nelder in Wedderburn. Danes se v številnih državah uporabljajo kot orodje za

določanje višine premij v zavarovalništvu, predvsem za neživljenjska zavarovanja (npr. za avtomobilsko zavarovanje). Pri premoženjskem zavarovanju se kot odvisna spremenljivka pojavlja škodna pogostost (angl. *claim frequency*), povprečna škoda (angl. *claim severity*), neto premija (angl. *pure premium*) ali škodni količnik (angl. *loss ratio*). Za te odvisne spremenljivke s pomočjo GLM ocenimo pričakovano vrednost le-teh. Po drugi strani lahko posplošene linearne modele uporabimo za ocenitev verjetnosti, da se bo zgodil določen dogodek. V takšnih primerih nas zanima, ali bo zavarovanec obnovil oz. prekinil svojo polico ali ne, ali prijavljen zahtevek vsebuje prevaro in podobno. Torej lahko modeliramo obnovljivost zavarovanj in analiziramo trajnost polic življenjskega zavarovanja [5].

Omenimo nekaj prednosti GLM v primerjavi z drugimi modeli. GLM so sestavljeni iz splošne statistične teorije, ki ima dobro razvite tehnike določanja standardnih napak, intervale zaupanja, testiranja, izbiranja modela in druge statistične značilnosti. Uporabni so na veliko področjih statistike in na voljo je precej orodij za delo s posplošenimi linearnimi modeli: SAS, R, GenStat, itd. [14].

Posplošeni linearni modeli so torej skupina naprednih statističnih metod, ki so posplošitev linearnih modelov in odpravljajo naslednji pomembni omejitvi teh modelov [14]:

1. *Verjetnostna porazdelitev.* Namesto predpostavke o normalni porazdelitvi podatkov, posplošeni linearni modeli predpostavljajo, da so Y_i porazdeljeni po poljubni porazdelitvi iz eksponentne družine porazdelitev, ki vsebujejo diskretne in zvezne porazdelitve, med drugimi tudi normalno, Poissonovo in gama porazdelitev.

2. *Model povprečja.* V linearnem modelu je pričakovana vrednost linearna funkcija neodvisnih spremenljivk, medtem ko je pri posplošenih linearnih modelih neka monotona transformacija pričakovane vrednosti linearno odvisna od neodvisnih spremenljivk.

GLM vsebuje naslednje komponente [3]:

1. Odvisne spremenljivke $Y_1, \ldots, Y_n$, ki so paroma neodvisne in imajo isto porazdelitev iz eksponentne družine porazdelitev.

2. *Linearni prediktor*: za naključne spremenljivke $Y_1, \ldots, Y_n$, je linearna komponenta (struktura) definirana kot $\eta_i = x_i\beta$, $i = 1, \ldots, n$, za vektor parametrov $\beta = (\beta_1, \ldots, \beta_k)^T$ in vektor pojasnjevalnih spremenljivk $x_i = (x_{1i}, \ldots, x_{ki})$, ki se nanaša na opazovan Y_i.

3. *Vezno funkcijo*: odvedljivo in strogo monotono funkcijo g, za katero velja $g_i(\mu_i) = \eta_i$, kjer je $\mu_i = E(Y_i)$.

V naslednjih poglavjih te komponente podrobneje predstavimo.

3.4 Družina eksponentnih porazdelitev

Družina eksponentnih porazdelitev posplošuje normalno porazdelitev, ki je uporabljena v linearnih modelih. Porazdelitev spada v eksponentno družino porazdelitev, če ima njena gostota verjetnosti (v primeru zvezne porazdelitve) oz. verjetnostna funkcija (v primeru diskretne porazdelitve) predpis

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + c(y_i,\phi,\omega_i)\right\}. \tag{3.1}$$

Parameter θ imenujemo kanonični parameter in je odvisen od opazovane enote i, medtem ko je disperzijski parameter $\phi > 0$ enak za vse i. Funkcija b je dvakrat zvezno odvedljiva z obrnljivim prvim odvodom [14]. Vrednost ω predstavlja utež, konstanto, ki je vnaprej določena. Pri zavarovalniških podatkih za utež pogosto uporabimo vrednost 1 (npr. za modeliranje števila škod), izpostavljenost (npr. za modeliranje škodne pogostosti) ali skupno število škod (npr. za modeliranje povprečne škode). Pri modeliranju škodne pogostosti se lahko opazovanja na primer nanašajo na enomesečno izpostavljenost ali na enoletno izpostavljenost. Z daljšim obdobjem izpostavljenosti zagotovimo več informacij in manj variabilnosti, kar je mogoče vključiti v model z opredelitvijo ω kot izpostavljenost vsakega opazovanja. Opazovanja z večjo izpostavljenostjo imajo nižjo varianco in bodo na model bolj vplivala, torej imajo takšna opazovanja večjo utež [1]. Funkcija c ni odvisna od θ in je v nekaterih primerih precej kompleksna, v teoriji posplošenih linearnih modelov pa nima posebnega pomena. Omenimo še naslednji tehnični omejitvi: utež je nenegativna (za vsak i je $\omega_i \geq 0$) in kanonični parameter θ_i za vsak i zavzame vrednosti v odprti množici (npr. $0 < \theta_i < 1$) [14].

Družina eksponentnih porazdelitev je zelo fleksibilna in se lahko uporablja za modeliranje zveznih (npr. normalna porazdelitev, gama porazdelitev), binarnih (npr. binomska porazdelitev) ali številnih podatkov (npr. Poissonova porazdelitev) [4].

3.4.1 Lastnosti družine eksponentnih porazdelitev

Izpeljimo matematično upanje in varianco za družino eksponentnih porazdelitev. Vsebina podpoglavja je povzeta po literaturi [14]. V nadaljevanju bomo zaradi poenostavitve zapisa opustili pisanje indeksa i. Najprej si poglejmo nekatere pojme in rezultate iz verjetnosti in statistike, ki nam bodo v pomoč pri izpeljavi.

Definicija 3.2 *Naj bo Y naključna spremenljivka. Če obstaja $E(|Y-a|^n)$, potem*

$$m_n(a) = E(|Y-a|^n) = \int_{\mathbb{R}} (y-a)^n dF$$

imenujemo moment reda n glede na a. Začetni n-ti moment $z_n = m_n(0)$ je moment glede na $a = 0$.

Definicija 3.3 *Eksponentna rodovna funkcija oz. rodovna funkcija momentov za slučajno spremenljivko Y je $M_Y(t) = E\left(e^{tY}\right)$, $t \in \mathbb{R}$, kjer matematično upanje obstaja. Natančneje,*

$$M_Y(t) = \int_{\mathbb{R}} e^{ty} f_Y(y) dy,$$

če je Y zvezna naključna spremenljivka in

$$M_Y(t) = \sum_{i=1}^{n} e^{ty_i} p_i,$$

če je Y diskretna naključna spremenljivka.

Trditev 3.4 *Naj bo Y slučajna spremenljivka. Če obstaja eksponentna rodovna funkcija $M_Y(t)$, potem velja $M_Y^{(n)}(t)\Big|_{t=0} = E(Y^n)$, kjer je $n \in \mathbb{N}$.*

Dokaz. Dokažimo trditev za primer zvezne naključne spremenljivke Y. Za diskretne spremenljivke je dokaz analogen.

Po definiciji je $M_Y(t) = \int_{\mathbb{R}} e^{ty} f_Y(y) dy$. Odvajamo n-krat po parametru t in dobimo enakost

$$M_Y^{(n)}(t) = \int_{\mathbb{R}} y^n e^{ty} f_Y(y) dy.$$

Sledi $M_Y^{(n)}(t)\Big|_{t=0} = M_Y^{(n)}(0) = \int_{\mathbb{R}} y^n f_Y(y) dy = E(Y^n)$. □

Definicija 3.5 *Kumulantna funkcija $\Psi_Y : \mathbb{R} \to \mathbb{R}$ je naravni logaritem eksponentne rodovne funkcije in je podana s predpisom $\Psi_Y(t) = \ln(M_Y(t))$.*

Definicija 3.6 *Kumulanta n-tega reda, označimo jo s κ_n, slučajne spremenljivke Y je n-ti odvod kumulantne funkcije v točki $t = 0$:*

$$\kappa_n = \Psi_Y^{(n)}(t)\Big|_{t=0}, \text{ kjer je } n \in \mathbb{N}. \tag{3.2}$$

Trditev 3.7 *Za kumulanto prvega reda velja* $\kappa_1 = E(Y)$ *in za kumulanto drugega reda velja* $\kappa_2 = \mathrm{Var}(Y)$.

Dokaz. Dokaz ločimo na dva dela.

a)

$$\kappa_1 = \Psi_Y^{'}(t)\Big|_{t=0} = \left(\ln\left(M_Y(t)\right)\right)^{'}\Big|_{t=0} = \frac{M_Y^{'}(t)}{M_Y(t)}\Bigg|_{t=0} = M_Y^{'}(0).$$

Sedaj upoštevamo še trditev 3.4: $M_Y^{'}(0) = E(Y)$.

b)

$$\begin{aligned}\kappa_2 = \ \Psi_Y^{''}(t)\Big|_{t=0} &= \left(\frac{M_Y^{'}(t)}{M_Y(t)}\right)^{'}\Bigg|_{t=0} = \frac{M_Y^{''}(t)M_Y(t) - M_Y^{'}(t)^2}{M_Y(t)^2}\Bigg|_{t=0} \\ &= M_Y^{''}(0) - M_Y^{'}(0)^2 = E(Y^2) - E(Y)^2 = \mathrm{Var}(Y).\end{aligned}$$

□

Sedaj se lahko posvetimo izpeljavi matematičnega upanja in variance za slučajne spremenljivke iz družine eksponentnih porazdelitev. Njihova eksponentna rodovna funkcija je (če matematično upanje obstaja za vsaj en $t \in \mathbb{R}$ v okolici 0) enaka [8]

$$\begin{aligned}M_Y(t) = E\left(e^{tY}\right) &= \int_{\mathbb{R}} e^{tY} f_Y(y;\theta,\phi)dy \\ &= \int_{\mathbb{R}} \exp\left\{\frac{y\left(\theta + t\frac{\phi}{\omega}\right) - b(\theta)}{\frac{\phi}{\omega}} + c(y,\phi,\omega)\right\} dy \\ &= \exp\left\{\frac{b\left(\theta + t\frac{\phi}{\omega}\right)}{\frac{\phi}{\omega}}\right\} \int_{\mathbb{R}} \exp\left\{\frac{y\left(\theta + t\frac{\phi}{\omega}\right) - b(\theta)}{\frac{\phi}{\omega}} + c(y,\phi,\omega)\right\} dy.\end{aligned}$$

Spomnimo se predpostavke, da mora kanonični parameter zavzeti vrednost v odprti množici. Sledi, da za vsaj en t v okolici 0 (za $|t| < \delta$ in za nek $\delta > 0$) velja, da je $\theta + t\frac{\phi}{\omega}$ v isti odprti množici. Na tem mestu za zadnji integral vidimo, da je integral gostote verjetnosti slučajne spremenljivke s parametrom $\theta + t\frac{\phi}{\omega}$ in uporabimo lastnost, da je integral gostote verjetnosti enak 1 ($\int_{\mathbb{R}} f(y;\theta + t\frac{\phi}{\omega},\phi)dy = 1$).

[8] V primeru diskretnih porazdelitev integracijo zamenjamo z vsotami.

Eksponentna rodovna funkcija je torej za $|t| < \delta$ enaka $M_Y(t) = \exp\left\{\frac{b\left(\theta+t\frac{\phi}{\omega}\right)}{\frac{\phi}{\omega}}\right\}$. Izraz logaritmiramo in dobimo kumulantno funkcijo $\Psi(t)$, ki obstaja za vsako porazdelitev z gostoto verjetnosti (3.1) in je za $|t| < \delta$ definirana s predpisom

$$\Psi(t) = \frac{b\left(\theta + t\frac{\phi}{\omega}\right) - b(\theta)}{\frac{\phi}{\omega}}. \tag{3.3}$$

Upoštevamo, da je funkcija b dvakrat odvedljiva (ena izmed predpostavk pri definiciji družine eksponentnih porazdelitev). Odvajamo funkcijo Ψ in dobimo

$$\Psi'(t) = b'\left(\theta + t\frac{\phi}{\omega}\right), \tag{3.4}$$

$$\Psi''(t) = b''\left(\theta + t\frac{\phi}{\omega}\right)\frac{\phi}{\omega}. \tag{3.5}$$

Če upoštevamo (3.4) ter enakost $E(Y) = \Psi'(0)$ sledi

$$\mu = E(Y) = b'(\theta). \tag{3.6}$$

Če upoštevamo (3.5) ter enakost $\mathrm{Var}(Y) = \Psi''(0)$ sledi

$$\mathrm{Var}(Y) = b''(\theta)\frac{\phi}{\omega}. \tag{3.7}$$

Izpeljali smo $\mu = b'(\theta)$. Iz predpostavke, da je odvod funkcije b obrnljiv, sledi, da je $\theta = b'^{-1}(\mu)$, kar vstavimo v $b''(\theta)$ in dobimo izraz $b''\left(b'^{-1}(\mu)\right)$, ki ga imenujemo variančna funkcija. Vidimo, da je odvisna od matematičnega upanja, zato jo lahko zapišemo kot $V(\mu)$. Sedaj lahko varianco zapišemo v naslednji obliki:

$$\mathrm{Var}(Y) = V(\mu)\frac{\phi}{\omega}. \tag{3.8}$$

3.4.2 Primer porazdelitve

Preverimo, da zelo znana (zvezna) normalna porazdelitev spada v družino eksponentnih porazdelitev. Normalna (Gaussova) porazdelitev $N\left(\mu_i, \sigma^2\right)$ ima gostoto verjetnosti podano s predpisom

$$f\left(y_i; \mu_i, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2}.$$

Predpostavimo torej, da je utež $\omega_i = 1$. Gostoto verjetnosti preoblikujemo na naslednji način:

$$f\left(y_i;\mu_i,\sigma^2\right) = e^{\ln(2\pi\sigma^2)^{-\frac{1}{2}}} e^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2} = e^{-\frac{1}{2}\ln(2\pi\sigma^2)-\frac{1}{2}\frac{y_i^2-2y_i\mu_i+\mu_i^2}{\sigma^2}}$$
$$= \exp\left\{\frac{y_i\mu_i - \frac{\mu_i^2}{2}}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right\}.$$

Opazimo, da normalna porazdelitev res spada v družino eksponentnih porazdelitev, saj velja $\theta_i = \mu_i$, $b(\theta_i) = \frac{\theta_i^2}{2}$, $\phi = \sigma^2$ in $c(y_i,\sigma^2,1) = -\frac{1}{2}\left(\frac{y_i^2}{\sigma_i^2} + \ln\left(2\pi\sigma^2\right)\right)$.

Veljata tudi naslednji povezavi, ki smo ju izpeljali v prejšnjem poglavju: $\mu = E(Y) = b'(\theta) = \theta$ in $\mathrm{Var}(Y) = b''(\theta)\phi = \sigma^2$. Variančna funkcija je $V(\mu) = 1$, posledično se varianca ne spreminja z matematičnim upanjem, odvisna spremenljivka je homoskedastična.

3.5 Vezna funkcija

Opomnimo, da linearni modeli predpostavljajo aditivni model za pričakovano vrednost. To lahko zapišemo kot [14]

$$\mu_i = \sum_{j=1}^{k} x_{ij}\beta_j; \quad i = 1,2,\ldots,n. \tag{3.9}$$

V GLM je leva stran zgornje enačbe posplošena na poljubno strogo monotono in odvedljivo funkcijo g. Ta funkcija se imenuje vezna oz. povezovalna funkcija (angl. *link function*), saj povezuje pričakovano vrednost z linearnim prediktorjem preko $g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j$. Posledično velja $\mu_i = g^{-1}(\eta_i)$.

Linearni modeli torej uporabljajo identiteto $g(\mu_i) = \mu_i$. Pri določanju premij za neživljenjska zavarovanja je najbolj pogosta logaritemska vez; $g(\mu_i) = \ln(\mu_i)$. Pogosto se uporablja tudi logit vez; $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$. Ta vez zagotavlja, da je pričakovana vrednost med 0 in 1. Več o tej vezni funkciji v poglavju o logistični regresiji 3.7.2.

V nekaterih primerih je učinek pojasnjevalne spremenljivke znan. Namesto ocenjevanja koeficienta β glede na to spremenljivko, vključimo znano informacijo o spremenljivki v model kot znani učinek. To je mogoče doseči z vključevanjem konstante, imenovane znan efekt ξ, v definicijo linearnega prediktorja η, ki ga lahko zapišemo kot $\eta = X\beta + \xi$. Na znan učinek lahko gledamo tudi kot na pojasnjevalno spremenljivko s koeficientom β enakim 1 [2]. Znan efekt pogosto predstavlja mero za izpostavljenost. Običajen primer uporabe znanega učinka je pri multiplikativnem modelu za določitev števila škod. Vsako opazovanje, ki se

nanaša na enomesečno izpostavljenost, bo očitno imelo nižje pričakovano število zahtevkov, pri čemer so vsi drugi dejavniki enaki, kot opazovanje v zvezi z enoletno izpostavljenostjo [1]. Tukaj je potrebno omeniti, da mora biti znan efekt istega tipa (v istem merilu) kot linearni prediktor. Kjer uporabljamo log vezno funkcijo, kot na primer pri modeliranju števila škod, uporabimo za znan efekt logaritem izpostavljenosti [17].

3.6 Struktura GLM

Z upoštevanjem do sedaj napisanega lahko posplošen linearni model zapišemo v obliki [1]

$$\mu_i = E(Y_i) = g^{-1}\left(\sum_{j=1}^{k} x_{ij}\beta_j + \xi_i\right) \quad \text{in} \quad \text{Var}(Y_i) = \frac{\phi V(\mu_i)}{\omega_i}, \tag{3.10}$$

kjer je

- Y_i vektor odvisnih spremenljivk,
- g vezna funkcija,
- x_{ij} elementi indikatorske matrike X (angl. *design matrix*),
- β_i vektor parametrov, ki jih ocenjujemo,
- ξ_i vektor znanega efekta (angl. *offset*),
- ϕ disperzijski parameter,
- V variančna funkcija in
- ω_i utež.

Vektor odvisnih spremenljivk, matriko pojasnjevalnih spremenljivk, uteži in znan efekt pridobimo iz podatkov. Obliko modela določimo s tem, da predpostavimo vezno funkcijo, variančno funkcijo ter disperzijski parameter [1].

3.7 Nekatere oblike GLM

$Y \sim$	Vez g	$E(Y) = \mu(\theta)$	$\mathrm{Var}(Y)$	$V(\mu)$	ϕ
$N(\mu, \sigma^2)$	identiteta	$\theta = \mu$	σ^2	1	σ^2
$\Gamma(\alpha, \lambda)$	inverz	$-\frac{1}{\theta} = \frac{\alpha}{\lambda}$	$\frac{1}{\alpha\theta^2} = \frac{\alpha}{\lambda^2}$	μ^2	α^{-1}
$Poiss(\lambda)$	logaritem	$e^{\theta} = \lambda$	$e^{\theta} = \lambda$	$e^{\theta} = \lambda = \mu$	1
$B(n, p)$	logit	$n\frac{e^{\theta}}{1+e^{\theta}} = np$	$np(1-p)$	$np(1-p)$	1
$IG(\mu, \lambda)$	$\frac{1}{\mu^2}$	$(-2\theta)^{-\frac{1}{2}} = \mu$	$\frac{\mu^3}{\lambda}$	μ^3	$\frac{1}{\lambda}$

Tabela 3.1: Nekatere porazdelitve iz družine eksponentnih porazdelitev

V tabeli 3.1, kjer smo predpostavili, da je utež $\omega = 1$, je prikazanih nekaj znanih porazdelitev iz družine eksponentnih porazdelitev. Vsaki porazdelitvi iz te družine pripada neka (naravna) vezna funkcija, za katero velja $g(.) = (b')^{-1}(.)$ (posledično velja $\theta = \eta$) in ji pravimo kanonična vezna funkcija (v tabeli 3.1: vez g). V splošnem velja [1]

$$\begin{aligned} f_{Y_i}(y_i; \theta_i, \phi) &= \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + c(y_i, \phi, \omega_i) \right\} \\ &= \exp\left\{ \frac{y_i(b')^{-1}\left(g^{-1}(\eta_i)\right) - b\left((b')^{-1}g^{-1}(\eta_i)\right)}{\frac{\phi}{\omega_i}} + c(y_i, \phi, \omega_i) \right\}. \end{aligned} \tag{3.11}$$

Če $\theta = \eta$, se zapis poenostavi na

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\eta_i - b(\eta_i)}{\frac{\phi}{\omega_i}} + c(y_i, \phi, \omega_i) \right\}. \tag{3.12}$$

Z izbiro takšne vezne funkcije si poenostavimo nekatere izračune (zgornji primer), vendar ni nujno, da je takšna izbira vedno najboljša. Na primer, McCullagh in Nelder predlagata inverzno vezno funkcijo (ki je tudi kanonična vez za gama porazdelitev) za modeliranje povprečne škode, vendar se v praksi pogosteje uporablja logaritemska vez [14].

V tabeli 3.2 je pregled nekaterih pogostih odvisnih spremenljivk v zavarovalništvu ter pripadajoča porazdelitev in parametri modela.

Pokažimo, da modeliranje števila škod, kjer je znan efekt enak logaritmu izpostavljenosti (utež je enaka 1) da enake rezultate kot modeliranje škodne pogostosti brez znanega efekta in z utežjo enako izpostavljenosti [1].

Naj naključna spremenljivka Y predstavlja število škod, ki se porazdeljuje po Poissonovi porazdelitvi. Uporabimo naravno vezno funkcijo za Poissonov model, ki je logaritem. Za

Y	škodna pogostost	število škod	povprečna škoda	verjetnost (npr. prekinitve)
$Y \sim$	Poisson	Poisson	gama	binomska
Vez $g(\mu)$	$\ln(\mu)$	$\ln(\mu)$	$\ln(\mu)$	$\ln\left(\frac{\mu}{1-\mu}\right)$
ϕ	1	1	ocenjen	1
$V(\mu)$	μ	μ	μ^2	$\mu(1-\mu)$ [9]
ξ	0	ln(izpostavljenost)	0	0
ω	izpostavljenost	1	število škod	1

Tabela 3.2: Pogosti modeli v zavarovalništvu [1]

znan efekt upoštevajmo logaritem izpostavljenosti. Model lahko zapišemo kot

$$\ln(E(Y|X)) = X\beta + \ln(izpostavljenost).$$

Iz tega sledi

$$X\beta = \ln(E(Y|X)) - \ln(izpostavljenost) = \ln\left(\frac{E(Y|X)}{izpostavljenost}\right).$$

Količnik $\frac{E(Y|X)}{izpostavljenost}$ predstavlja škodno pogostost. Model za število škod z znanim efektom smo torej preoblikovali v model za škodno pogostost brez znanega efekta, ampak z utežjo enako izpostavljenosti.

3.7.1 Primer Poissonovega modela

Za ilustracijo si poglejmo uporabo Poissonovega modela [3]. Če privzamemo, da je smrt zaradi nenalezljive bolezni neodvisni dogodek, potem lahko število smrti Y modeliramo s Poissonovo porazdelitev z verjetnostno funkcijo, ki ima predpis $f(y;\mu) = \frac{\mu^y e^{-\mu}}{y!}$, kjer je $y \in \{0, 1, 2, 3, \dots\}$, in $\mu = E(Y)$ je pričakovano število smrti v določen obdobju, na primer v enem letu. Parameter μ bo odvisen od velikosti populacije, obdobja opazovanja in številnih značilnosti populacije (npr. starost, spol, zdravstveno stanje). Lahko je modeliran na primer z $\mu = n\lambda(X\beta)$, kjer je n velikost populacije in $\lambda(X\beta)$ stopnja umrljivosti na 100.000 ljudi na leto (kar je odvisno od populacijskih značilnosti, ki jih opisuje linearna komponenta $X\beta$).

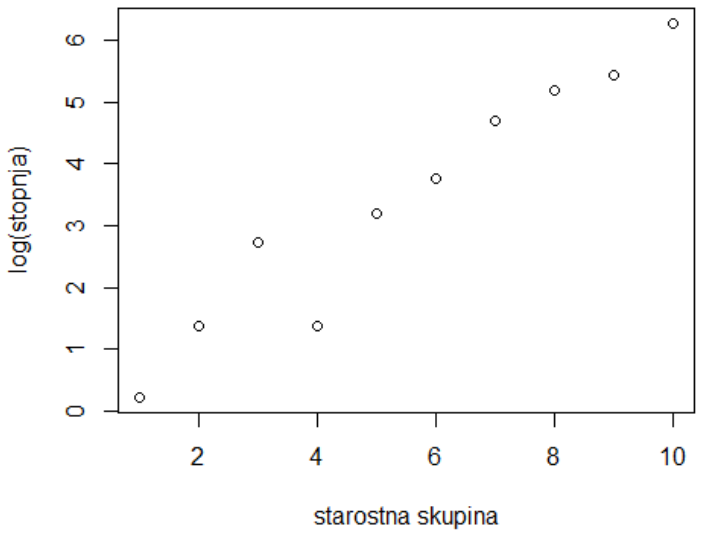
Spremembo umrljivosti s starostjo lahko modeliramo tako, da naključne spremenljivke $Y_1, Y_2, \dots, Y_n$ upoštevamo kot število smrti v zaporednih starostnih skupinah. Tabela 3.3

[9] Dan izraz je za primer, kjer je število poizkusov enako 1. V primeru, da je število poizkusov enako t, je izraz enak $\frac{\mu(t-\mu)}{t}$.

prikazuje podatke za leto 2017 glede na starost za smrt moških zaradi bolezni živčevja v Sloveniji.

Starostna skupina	Starostna skupina (leta)	Število smrti	Velikost populacije	Stopnja na 100.000 moških na leto
1	40–44	1	79484	1,3
2	45–49	3	75371	4,0
3	50–54	12	78796	15,2
4	54–59	3	74773	4,0
5	60–64	18	73377	24,5
6	65–69	25	58547	42,7
7	70–74	42	38793	108,3
8	75–79	57	32199	177,0
9	80–84	47	20774	226,2
10	85+	68	12911	526,7

Tabela 3.3: Število smrti za bolezen živčevja in število moške populacije glede na starostno skupino v Sloveniji leta 2017 (Vir: Statistični urad Republike Slovenije)



Slika 3.1: Stopnja umrljivosti na 100.000 moških.

Slika 3.1 prikazuje, kako se stopnja umrljivosti $\frac{y_i}{n_i} \cdot 100.000$ povečuje s starostjo. Na navpični osi je uporabljen logaritem stopnje umrljivosti. Iz grafa je razvidno, da je logaritem stopnje

umrljivosti približno linearno odvisen od starosti. To sugerira, da je odnos med $\frac{y_i}{n_i}$ in starostno skupino i približno eksponenten. Zato je možen model:

$$E(Y_i) = \mu_i = n_i e^{\theta i}, \quad Y_i \sim Poiss(\mu_i),$$

kjer je $i = 1$ za starostno skupino 40–44, $i = 2$ za starostno skupino 45–49, ..., $i = 10$ za starostno skupino 85+. To lahko zapišemo kot posplošen linearni model z uporabo logaritemske vezne funkcije

$$g(\mu_i) = \ln(\mu_i) = \ln(n_i) + \theta i$$

in linearne komponente $x_i\beta$, kjer je $x_i = [\ln(n_i) \quad i]$ in $\beta = [1 \quad \theta]^T$. Opazimo, da ima velikost populacije n vlogo znanega efekta.

3.7.2 Logistična regresija

Naj bo Y binarna odvisna spremenljivka, ki zavzame dve vrednosti; 1 in 0. Naj bo $P(Y = 1) = p$ in $P(Y = 0) = 1 - p$. Potem velja $Y \sim B(1, p)$ in $E(Y) = \mu = p$, $\mathrm{Var}(Y) = p(1 - p)$. Takšni porazdelitvi pravimo Bernoullijeva porazdelitev in je poseben primer Binomske porazdelitve, kjer je število neodvisnih poskusov enako 1.

Odvisna spremenljivka, ki je porazdeljena po Bernoullijevi porazdelitvi, in logit vezna funkcija definirata logistično regresijo. O logistični regresiji torej govorimo kadar imamo opravka z odvisno spremenljivko, ki je kategorična in nas zanima ali se bo določen dogodek zgodil ali ne.

Izraz $\frac{p}{1-p}$ je razmerje med verjetnostjo, da se dogodek zgodi in verjetnostjo, da se dogodek ne zgodi in se imenuje obeti. Ekvivalentno, obeti povedo kolikokrat je verjetnost pojava dogodka večja od verjetnosti, da se dogodek ne zgodi.

Logit vez $g(\mu) = \ln \frac{p}{1-p} = X\beta$ lahko interpretiramo kot logaritem obetov. Iz tega sledi $p = \frac{e^{X\beta}}{1+e^{X\beta}}$. Logit vezna funkcija zagotavlja, da verjetnost p leži na intervalu $(0, 1)$ za vse vrednosti β in X [2].

Razlaga regresijskih koeficientov

Najprej si poglejmo interpretacijo regresijskih koeficientov na primeru ene zvezne pojasnjevalne spremenljivke x. Vezna funkcija je v tem primeru enaka $\ln \frac{p}{1-p} = \beta_1 + \beta_2 x$. Če na zgornjo enačbo delujemo s funkcijo e (antilogaritmiramo), dobimo, da so obeti enaki $e^{\beta_1+\beta_2 x}$. Če se pojasnjevalna spremenljivka x poveča za eno enoto, potem so obeti enaki

$e^{\beta_1+\beta_2(x+1)} = e^{\beta_2}e^{\beta_1+\beta_2 x}$. Z drugimi besedami, povečanje pojasnjevalne spremenljivke za eno enoto pomeni množenje obetov z e^{β_2}. Če to posplošimo na model z več zveznimi pojasnjevalnimi spremenljivkami, potem povečanje j-te pojasnjevalne spremenljivke za eno enoto povzroči množenje obetov z e^{β_j}, pri čemer vse ostale pojasnjevalne spremenljivke ostanejo nespremenjene.

Sedaj predpostavimo, da je pojasnjevalna spremenljivka x kategorična, ki zavzame r kategorij. Za bazno kategorijo izberemo r-to kategorijo. Vezno funkcijo zapišemo kot

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{r-1} x_{r-1},$$

kjer so x_j indikatorske spremenljivke. Spremenljivka x_j je enaka 1, če kategorična pojasnjevalna spremenljivka x zavzame vrednost j-te kategorije ($j = 1, 2, \ldots, r-1$) in 0 sicer. Če x zavzame vrednost bazne kategorije, potem so obeti enaki e^{β_0}. Če x zavzame vrednost kategorije j, ki ni bazna kategorija, so obeti enaki $e^{\beta_0}e^{\beta_j}$. Učinek spremenljivke x, ki se nahaja v kategoriji j v primerjavi z bazno kategorijo, je množenje obetov e^{β_0} s faktorjem e^{β_j} [2].

3.8 Ocenjevanje regresijskih koeficientov in parametra ϕ

Do zdaj smo spoznali, kako so definirani posplošeni linearni modeli, sedaj pa se posvetimo najpomembnejšemu koraku, ocenjevanju regresijskih koeficientov in parametrov. Najpogostejša metoda ocenjevanja regresijskih koeficientov β je metoda največjega verjetja, ki pri linearni regresiji da enake rezultate kot metoda najmanjših kvadratov [3].

3.8.1 Metoda največjega verjetja

Koeficienti β so določeni z metodo največjega verjetja (angl. *maximum likelihood*) Denimo, da imamo v vzorcu n neodvisnih opazovanj in smo v model vključili k pojasnjevalnih spremenljivk. Potem je verjetje definirano kot

$$L = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta_i, \phi)$$

in posledično je logaritem verjetja (log-verjetje) enak

$$\ell = \ln L = \sum_{i=1}^{n} \ln \left(f_{Y_i}(y_i; \theta_i, \phi) \right).$$

Za nadaljnje izračune bomo uporabili logaritem verjetja, saj je delo z njim lažje, maksimiranje funkcije verjetja pa je ekvivalentno maksimiranju log-verjetja. Cilj je torej določiti takšne koeficiente β, da bo funkcija ℓ maksimalna.

Log-verjetje za porazdelitev z enačbo (3.1) kot funkcija θ je

$$\ell(\theta; \phi, y) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + c(y_i, \phi, \omega_i) \right), \tag{3.13}$$

kar lahko zapišemo kot

$$\ell(\theta; \phi, y) = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \left(y_i\theta_i - b(\theta_i) \right) + \sum_{i=1}^{n} c(y_i, \phi, \omega_i). \tag{3.14}$$

Disperzijski parameter ϕ ne vpliva na maksimiranje ℓ glede na θ in ga zato na tem mestu lahko ignoriramo (ocenjevanju parametra ϕ se bomo posvetili kasneje). Verjetje kot funkcijo od β izpeljemo s pomočjo inverzne funkcije od $\mu_i = b'(\theta_i)$ in vezne funkcije $g(\mu_i) = \eta_i = \sum_j x_{ij}\beta_j$.

Log verjetje je največje, če za vsak j izračunamo parcialni odvod funkcije ℓ glede na β_j in ga enačimo z 0: $\frac{\partial \ell}{\partial \beta_j} = 0$.

Za ocenitev parametrov β_j uporabimo verižno pravilo za odvajanje. Odvod ℓ glede na β_j je

$$\begin{aligned} 0 = \frac{\partial \ell}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial \ell}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \left(\omega_i y_i - \omega_i b'(\theta_i) \right) \frac{\partial \theta_i}{\partial \beta_j} \\ &= \frac{1}{\phi} \sum_{i=1}^{n} (\omega_i y_i - \omega_i b'(\theta_i)) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \end{aligned}$$

Sedaj določimo parcialne odvode v zgornji formuli. Če enakost $\mu_i = b'(\theta_i)$ odvajamo po parametru θ_i, dobimo zvezo

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i). \tag{3.15}$$

Združimo enakost, ki je izpeljana iz enačbe (3.6) v poglavju 3.4.1, $V(\mu) = b''(\theta_i)$, ter zvezo (3.15) in dobimo naslednjo enakost: $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{V(\mu_i)}$.

Poleg tega velja $\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} = \frac{1}{g'(\mu)}$ in $\eta_i = \sum_j x_{ij}\beta_j \Rightarrow \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$.

Z upoštevanjem zgornjih zvez dobimo sistem n enačb:

$$0 = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\omega_i(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)} x_{ij}, \text{ kjer je } j = 1, 2, \ldots k. \tag{3.16}$$

Sistem (3.16) lahko v matrični enačbi zapišemo kot

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

kjer je W diagonalna matrika z diagonalnimi elementi

$$\tilde{\omega_i} = \frac{\omega_i}{V(\mu_i) g'(\mu_i)}, \quad W = \mathrm{diag}(\tilde{\omega_i}; i = 1, \ldots, n).$$

Direktno iskanje rešitve tega sistema je zapleteno. Zaradi velike količine podatkov, s katerimi ponavadi razpolagamo, eksplicitno ocenjevanje regresijskih koeficientov ni smiselno in se raje poslužujemo iterativnih numeričnih metod. Za rešitev tega sistema pogosto uporabimo Newtonovo metodo [1].

V splošnem z Newtonovo metodo iščemo rešitev enačbe $f(x) = 0$, kjer mora biti f odvedljiva funkcija, z začetno vrednostjo x_1 in iteracijo $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, za vsak $n \in \mathbb{N}$.

V našem primeru je iteracijski korak Newtonove metode:

$$\beta_{n+1} = \beta_n - \left(H^{-1}\right)_n \left(X^T W_n y - X^T W_n \mu_n\right),$$

kjer je H Hessejeva matrika in vsebuje parcialne odvode drugega reda log-verjetja. Na vsakem koraku moramo torej izračunati H, W in μ [14].

Omenimo še dejstvo, da so ocene regresijskih koeficientov asimptotično normalno porazdeljene, $\hat{\beta} \sim N\left(\beta, \phi \left(X^T W X\right)^{-1}\right)$. Skica dokaza je podana v literaturi [14], na strani 51.

3.8.2 Ocenjevanje parametra ϕ

V nekaterih primerih (npr. pri Poissonovi porazdelitvi) je disperzijski parameter ϕ enak 1 in ga torej ni potrebno ocenjevati. V splošnem ϕ ni vnaprej poznan in v takšnem primeru ga moramo oceniti s pomočjo podatkov. Kot smo videli zgoraj, ocenjen disperzijski parameter ne vpliva na ocenitev regresijskih koeficientov β. Potrebovali pa ga bomo za določitev določenih statistik (npr. F-statistike) [1]. Disperzijski parameter lahko določimo s pomočjo metode največjega verjetja. Slabost tega pristopa je ta, da ni mogoče izpeljati eksplicitne formule za ϕ in je lahko takšno ocenjevanje dolgotrajno. Metoda največjega verjetja namreč vključuje iterativno reševanje enačbe $\frac{\partial \ell}{\partial \phi} = 0$, ki pa je drugačna za vsako porazdelitev odvisne spremenljivke [2].

Kot alternativo lahko uporabimo Pearsonovo χ^2-statistiko, ki je klasično merilo za ocenjevanje primernosti statističnega modela [14]. Za posplošen linearni model je ta statistika

definirana kot [10]

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\mathrm{Var}(Y_i)} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{\omega_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

kjer je $\hat{\mu}_i$ ocenjena vrednost matematičnega upanja in y_i vrednost odvisne spremenljivke pri i-ti opazovani enoti. Iz statistike je znano, da se χ^2 porazdeljuje približno kot $\chi^2(n-k)$ in je zato matematično upanje porazdelitve χ^2 približno enako $n-k$. Tako dobimo cenilko za ϕ (imenovano tudi momentno cenilko):

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

Tretji način ocenjevanja disperzijskega parametra je s pomočjo deviance D, ki jo bomo definirali kasneje. Cenilka za ϕ je potem enaka

$$\hat{\phi}_D = \frac{D}{n-k}.$$

3.9 Testiranje signifikantnosti pojasnjevalnih spremenljivk

Pomembno vprašanje za vsak ocenjen regresijski koeficient je, ali je le-ta dovolj blizu »resnične« vrednosti koeficienta oz. še bolj pomembno vprašanje, ali ima pojasnjevalna spremenljivka kakšen vpliv na odvisno spremenljivko. Potrebno je preveriti, ali smo od nič različno vrednost dobili zato, ker je tudi njegova dejanska vrednost različna od nič, ali smo to vrednost dobili po naključju.

Ničelna in alternativna hipoteza sta

$$H_0 : \beta_j = 0 \quad \text{in} \quad H_1 : \beta_j \neq 0, \quad j = 1, \ldots, k.$$

Spomnimo, da se cenilka $b = (b_1, b_2, \ldots, b_k)$ porazdeljuje asimptotično normalno. Uporabimo t-statistiko za preizkušanje domnev o vrednosti regresijskih koeficientov. Preverjamo torej, ali je vrednost regresijskih koeficientov statistično značilna. t-statistika se izračuna kot

$$t_j = \frac{b_j - \beta_j}{se(b_j)},$$

kjer je $se(b_j)$ standardna napaka ocene regresijskega koeficienta b_j (definirana kot kvadratni

[10] Upoštevamo, da velja $\mathrm{Var}(Y_i) = \frac{\phi V(\hat{\mu}_i)}{\omega_i}$.

koren diagonalnih elementov kovariančne matrike). Vrednost t_j je porazdeljena po Studentovi t-porazdelitvi z $(n-k)$ prostostnimi stopnjami.

Če velja $|t_j| > t_c^{\alpha/2}(n-k)$ [11], potem ničelno domnevo zavrnemo in sprejmemo sklep, da je regresijski koeficient β_j statistično značilno različen od nič pri stopnji značilnosti preizkusa α.

Statistični računalniški programi izračunavajo tudi t.i. vzorčno signifikanco ali p-vrednost (prava oz. natančna stopnja značilnosti). H_0 zavrnemo pri vseh stopnjah značilnosti, za katere velja $\alpha \geq p$ [15].

Pomembno orodje za presojo natančnosti ocenjenih regresijskih koeficientov so tudi intervali zaupanja. Intervali zaupanja dajo območje verjetnih vrednosti za regresijski koeficient β_j.

$100(1-\alpha)$-odstotni interval zaupanja za vrednosti t-statistike je enak

$$P\left(-t_{\alpha/2} \leq t = \frac{b_j - \beta_j}{se(b_j)} \leq t_{\alpha/2}\right) = 1 - \alpha,$$

kar lahko preoblikujemo v zapis

$$P\left(b_j - t_{\alpha/2} se(b_j) \leq \beta_j \leq b_j + t_{\alpha/2} se(b_j)\right) = 1 - \alpha.$$

Zgornjo enačbo si razlagamo na naslednji način: »verjetnost, da na podlagi vzorčnih podatkov oblikujemo interval, ki bo vseboval β_j, je enaka $(1-\alpha)$« [15]. Interval zaupanja je naključen; na podlagi vzorčnih podatkov izračunamo le enega izmed vseh možnih intervalov. Med temi intervali je $100(1-\alpha)\,\%$ takšnih, ki vključujejo dejansko vrednost regresijskega koeficienta, $100\alpha\,\%$ pa jih ne vključuje.

Če ocenjen regresijski koeficient ni statistično značilen, bo interval zaupanja vseboval število nič. V tem primeru lahko rečemo, da na podlagi vzorčnih podatkov ugotavljamo, da pojasnjevalna spremenljivka nima pomembnega vpliva na odvisno spremenljivko in jo izključimo iz modela. Tukaj je potrebno poudariti, da statistično značilna spremenljivka ne pomeni nujno, da ta predstavlja smiselno odvisnost. Npr. na podlagi regresijskega modela ugotavljamo, da barva oči statistično značilno vpliva na smrtnost, kar ne pomeni, da je to tudi smiselna ugotovitev. S pomočjo intervalov zaupanja presojamo tudi o združitvi kategorij spremenljivke, ki niso statistično značilne. Če interval zaupanja določene kategorije posamezne pojasnjevalne spremenljivke vsebuje interval zaupanja druge kategorije, ti dve kategoriji združimo.

[11] $t_c^{\alpha/2}(n-k)$ je kritična vrednost t-statistike pri stopnji značilnosti α in $n-k$ prostostnimi stopnjami.

3.10 Mere primernosti regresijskega modela

3.10.1 Devianca

Eden izmed načinov ocenjevanja primernosti modela je primerjava z bolj splošnim modelom, ki ima maksimalno število pojasnjevalnih spremenljivk, ki jih je mogoče oceniti. Takšen model se imenuje nasičen model (angl. *saturated model*). To je posplošen linearni model z enako porazdelitvijo in vezno funkcijo kot model, ki ga ocenjujemo. Če imamo n opazovanih enot (meritev) Y_i; $i = 1, \ldots, n$, ki imajo različne vrednosti za linearno komponento, potem je nasičen model določen z n pojasnjevalnimi spremenljivkami. Takšen model imenujemo tudi maksimalen oz. polni model. Le-ta se popolnoma prilega danim podatkom in bi natančno ≫napovedal≪ rezultat za zgodovinske podatke - vsaka ocenjena vrednost je enaka dejanski vrednosti odvisne spremenljivke. Takšen model pa je seveda zgolj hipotetičen in neuporaben za napovedovanje, vseeno pa dobro služi za zgornjo mejo log-verjetja za dane podatke [5].

Naj $\ell(\hat{\mu})$ označuje log-verjetje za model, ki ga ocenjujemo in naj $\ell(y)$ označuje log-verjetje za polni model, ki bo seveda večji od vseh drugih log-verjetij za te opazovane enote z isto porazdelitvijo in vezno funkcijo, saj predstavlja najbolj popoln opis podatkov.

Skalirana devianca D^* je definirana kot mera za razdaljo med nasičenim in dejanskim (ocenjenim) modelom [14] [12]:

$$D^* = D^*(y, \hat{\mu}) = 2(\ell(y) - \ell(\hat{\mu})).$$

Naj h označuje inverzno funkcijo funkcije b' v enakosti $\mu_i = b'(\theta_i)$. Torej velja $\theta_i = h(\mu_i)$. Če upoštevamo še enakost za log-verjetje (3.14), lahko skalirano devianco zapišemo kot

$$D^* = \frac{2}{\phi} \sum_i \omega_i (y_i h(y_i) - b(h(y_i)) - y_i h(\hat{\mu_i}) + b(h(\hat{\mu_i}))), \tag{3.17}$$

Z množenjem zgornje enačbe s parametrom ϕ dobimo (neskalirano) devianco, ki je neodvisna od disperzijskega parametra:

$$D = \phi D^*. \tag{3.18}$$

Če definiramo

$$d(y, \mu) = 2(yh(y) - b(h(y)) - yh(\mu) + b(h(\mu))), \tag{3.19}$$

[12] Pri obeh log-verjetjih je uporabljena ista vrednost za ϕ.

potem iz (3.17) in (3.18) sledi, da lahko devianco zapišemo kot $D = \sum_{i=1}^{n} \omega_i d(y_i, \hat{\mu}_i)$.

Devianco lahko torej interpretiramo kot uteženo vsoto razlike med ocenjenim matematičnim upanjem $\hat{\mu}_i$ in opazovano vrednostjo y_i.

Če je $\ell(\hat{\mu})$ blizu vrednosti $\ell(y)$, potem se pričakuje, da model dobro opisuje dano situacijo. V nasprotnem primeru pa velika vrednost deviance pomeni, da imamo opravka z modelom, ki slabo ustreza podatkom [2].

3.10.2 Primerjava gnezdenih modelov

Devianca se uporablja pri gnezdenih modelih. Gnezdena modela sta modela z enako verjetnostno porazdelitvijo in enako vezno funkcijo, kjer je linearna komponenta preprostejšega modela M_o poseben primer linearne komponente splošnejšega modela M_s. Povedano drugače, en model vsebuje vse pojasnjevalne spremenljivke drugega modela in eno ali več dodatnih pojasnjevalnih spremenljivk. Na tej točki nas zanima, ali je model z dodatnimi pojasnjevalnimi spremenljivkami signifikantno boljši [3].

V primeru, da je disperzijski parameter znan, kot npr. pri Poissonovi porazdelitvi, lahko uporabimo χ^2-test. Ničelna hipoteza je naslednja: dodatne spremenljivke bistveno ne prispevajo k pojasnitvi modela. Izračunamo razliko med skaliranima deviancama modelov, ki se porazdeljuje po χ^2-porazdelitvi z m prostostnimi stopnjami (m je število dodanih pojasnjevalnih spremenljivk):

$$\Delta D_o^* - D_s^* = 2(\ell(y) - \ell_o(\hat{\mu})) - 2(\ell(y) - \ell_s(\hat{\mu})) = 2(\ell_s(\hat{\mu}) - \ell_o(\hat{\mu})) \sim \chi^2(m).$$

Če izračunana vrednost χ^2-statistike pade v kritično območje pri dani stopnji značilnosti α, zavrnemo ničelno hipotezo in sprejmemo sklep, da dodatne spremenljivke signifikantno zmanjšajo devianco [1].

F-test uporabimo, ko je disperzijski parameter ocenjen (npr. pri gama porazdelitvi). Ničelna hipoteza je enaka kot pri χ^2-testu. F-statistiko izračunamo kot

$$F = \frac{D_o - D_s}{m\hat{\phi}_o} \sim F(m, n - k_o), \tag{3.20}$$

kjer je D_o devianca manjšega modela, D_s devianca modela z dodanimi m pojasnjevalnimi spremenljivkami, n velikost vzorca, k_o število pojasnjevalnih spremenljivk manjšega modela ter $\hat{\phi}_o$ ocenjen disperzijski parameter osnovnega modela [5].

Če izračunana vrednost F-statistike pade v kritično območje ($F > F_c^{\alpha}(m - k_o)$) [13] pri dani stopnji značilnosti α, zavrnemo ničelno hipotezo in sprejmemo sklep, da model M_s predstavlja signifikantno boljši opis podatkov.

3.10.3 Akaikov informacijski kriterij

Devianca ni uporabna za negnezdene modele, kjer en model ne vsebuje podmnožice pojasnjevalnih spremenljivk drugega modela. Dodajanje spremenljivk vedno zmanjša devianco in lahko privede do prevelikega prileganja podatkom (angl. *over-fitting*). Model, ki bi imel več pojasnjevalnih spremenljivk, bi potem bil vedno boljši, kar pa ni nujno res.

Če želimo primerjati negnezdene modele, uporabimo Akaikov informacijski kriterij, ki temelji na devianci, a kaznuje dodajanje pojasnjevalnih spremenljivk v model. Definiran je kot

$$AIC = -2\ell + 2k,$$

kjer je ℓ log-verjetje ocenjenega modela in k število pojasnjevalnih spremenljivk [5].

Podobno kot pri popravljenem determinacijskem koeficientu R^2 za prostostne stopnje pri linearni regresiji, [14] je namen AIC preprečiti, da bi vključili nepomembne pojasnjevalne spremenljivke. Vendar za razliko od popravljenega R^2 za prostostne stopnje, sama vrednost AIC ne pove veliko. Uporaben je zgolj za primerjavo različnih modelov. Bolšji model je model z nižjo vrednostjo AIC [8].

3.10.4 Ostanki

Z analizo ostankov določimo, kako dobro ocenjen model ustreza podatkom. Analiziramo torej odstopanja posameznih podatkovnih točk od njihovih napovedanih vrednosti. Za ocenitev kvalitete modela in identifikacijo osamelcev si pogosto pomagamo s histogramom porazdelitve ostankov regresijskega modela in grafikonom Q-Q. Poglejmo si enega izmed tipov ostankov za določanje, kako se ocenjene vrednosti razlikujejo od dejanskih vrednosti za vsako opazovanje.

Ostanki pri linearnem modelu so definirani kot $e_i = y_i - \hat{\mu}_i$, vendar pri posplošenih linearnih modelih ne povedo veliko. Za GLM ti ostanki niso niti normalno porazdeljeni, niti nimajo konstantne variance, za katerokoli odvisno spremenljivko, drugačno od normalne [2].

[13] $F_c^{\alpha}(m - k_o)$ je kritična vrednost F-statistike pri stopnji značilnosti α in $m - k_o$ prostostnimi stopnjami.

[14] Determinacijski koeficient R^2 pove, koliko odstotkov variance odvisne spremenljivke smo pojasnili z ocenjenim modelom. Več o determinacijskem koeficentu pri linearni regresiji najdemo v knjigi Osnove ekonometrije na strani 114.

Ostanki deviance

Ostanki deviance se uporabljajo za preverjanje ustreznosti izbrane porazdelitve odvisne spremenljivke. V dobrem modelu pričakujemo, da bodo ostanki deviance (angl. *deviance residuals*) imeli naslednje lastnosti [5]:

1. so naključni, torej ne sledijo kakšnemu vzorcu;
2. so normalno porazdeljeni s konstantno varianco (homoskedastični). Če opazimo drugačno porazdelitev ostankov ali heteroskedastičnost, je to sum, da je izbrana porazdelitev nepravilna.

Ostanki deviance so definirani kot

$$r_{D_i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{\omega_i d(y_i, \hat{\mu}_i)},$$

kjer je d funkcija, ki smo jo definirali v poglavju 3.10.1 in

$$\text{sign}(y_i - \hat{\mu}_i) = \begin{cases} 1; & y_i > \hat{\mu}_i \\ -1; & y_i < \hat{\mu}_i \end{cases}.$$

Očitno velja tudi naslednja povezava: $\sum_{i=1}^{n} r_{D_i}^2 = \theta D^* = D$. Z drugimi besedami, devianca je vsota kvadratov ostankov z ustreznim predznakom [14].

Opomba 3.8 *Za diskretne porazdelitve so ostanki deviance manj uporabni kot določitev primernosti porazdelitev, saj je velika verjetnost, da se ostanki deviance ne bodo porazdelili po normalni porazdelitvi. Veliko število opazovanj ima namreč enako vrednost odvisne spremenljivke in to povzroči, da se ostanki združijo v tesne skupine [5].*

3.11 Prerazpršenost

Če je opažena varianca večja od variance teoretičnega modela, pravimo, da je prišlo do prerazpršenosti (prekomerne disperzije). McCullagh in Nelder v [10] navajata, da je prerazpršenost prej pravilo kot izjema. Nasprotno, premajhna razpršenost pomeni, da podatki manj variirajo kot pričakovano, kar pa je v praksi redkost. Na prerazpršenost pogosto naletimo pri modeliranju številnih podatkov, kjer uporabljamo binomsko (za $n > 1$) ali Poissonovo porazdelitev (npr. modeliranje škodne pogostosti).

Prerazpršenost lahko povzročijo sistematične pomanjkljivosti modela. Do tega lahko pride zaradi izpustitve pomembne pojasnjevalne spremenljivke v modelu, zaradi nepravilno izbrane funkcijske oblike modela (npr. izbira linearne oblike namesto nelinearne), napačne izbire porazdelitve odvisne spremenljivke ali vezne funkcije. Do prerazpršenosti lahko pride tudi, če je model ustrezen. V tem primeru ta pojav povzroča nepojasnjeno naključno variiranje.

Če je v naboru podatkov prisotna prevelika razpršenost, bodo ocenjene standardne napake pristranske oz. natančneje, prenizke ter t-statistike visoke. Posledično pridemo do napačnih ugotovitev glede signifikantnosti regresijskih koeficientov (napaka I. vrste [15]).

Za ugotavljanje, ali je v modelu prisotna prekomerna varianca, lahko uporabimo pravilo zidarskega palca (angl. *the rule of thumb*). Ta pravi, da mora devianca deljena s prostostnimi stopnjami biti približno 1. Vrednosti, ki so večje kot ena, sugerirajo na prekomerno varianco. V nasprotnem primeru lahko ocenimo, da gre za podrazpršenost. Na tem mestu moramo poudariti, da je to pravilo asimptotsko. Uporabimo ga lahko zgolj za velike vrednosti matematičnega upanja ($\mu_i > 5$) pri Poissonovi porazdelitvi in za velike $n_i p_i$ pri binomski porazdelitvi [16].

Pri Poissonovi porazdelitvi predpostavimo, da je matematično upanje enako varianci, vendar je v praksi varianca pogosto večja. Formalno to preverimo s testom prerazpršenosti. Ideja testa je naslednja. Za Poissonov model velja $E(Y) = \mu = \mathrm{Var}(Y)$. To je tudi hipoteza, ki jo želimo preveriti. Alternativna hipoteza je $\mathrm{Var}(Y) = \mu + cf(\mu)$, kjer konstanta $c < 0$ implicira podrazpršenost in $c > 0$ pomeni prerazpršenost. Funkcija f je neka monotona funkcija (pogosto kvadratna ali linearna, ki je privzeta). Preverjamo torej hipotezo $H_0 : c = 0$ proti alternativni $H_1 : c \neq 0$. Testna statistika, ki je uporabljena je t-statistika.

Za določitev, kdaj je prisotna prerazpšenost problematična in kdaj je potrebno ukrepati, ni jasno določene meje. V primeru ugotovitve, da je v modelu prisotna prekomerna varianca, ustrezno ukrepamo.

Eden izmed ukrepov je uporaba kvazipoissonovega modela namesto Poissonovega. Kvazipossonov model je podoben kot Poissonov, pri čemer je glavna razlika disperzijski parameter. Le-ta omogoča, da parameter ni nujno omejen na vrednost 1, ampak lahko zavzame katerokoli pozitivno vrednost. Modela dasta enake ocene regresijskih koeficientov in s tem enake napovedi; drugačne pa so ocene standardnih napak (te so pomnožene s korenom disperzijskega parametra)[5].

[15] Zavrnemo $H_0 : \beta = 0$ in sprejmemo sklep, da je regresijski koeficient statistično značilen, čeprav v resnici ni.

3.12 Interakcije

Do sedaj smo predpostavili, da ima vsaka posamezna pojasnjevalna spremenljivka individualni vpliv na odvisno spremenljivko. Vendar lahko imata dve ali več pojasnjevalnih spremenljivk medsebojni vpliv na odvisno spremenljivko. Povedano drugače, vpliv ene pojasnjevalne spremenljivke na odvisno spremenljivko je lahko odvisen od vrednosti druge pojasnjevalne spremenljivke in obratno. Tak odnos se imenuje interakcija [5].

Pri modeliranju škodne pogostosti se pogosto upošteva interakcija med spolom in starostjo. Mladi moški so namreč bolj nagnjeni k povzročitvi prometne nesreče kot mlade ženske, medtem ko pri starejših, razlike med spoloma v vozniškem obnašanju niso očitne.

Poglavje 4

Uporaba GLM na primeru avtomobilskega zavarovanja

Neto premijo za avtomobilsko zavarovanje lahko modeliramo na dva načina. Zgradimo lahko en model, ki ima kot odvisno spremenljivko neto premijo. To lahko dosežemo s pomočjo Tweedie porazdelitve, ki pa je ne bomo podrobneje obravnavali. Drugi način, kateri bo uporabljen v nalogi, posebej obravnava škodno pogostost in povprečno škodo. Neto premija je produkt teh dveh rezultatov. Modeliranje na tak način ima kar nekaj prednosti; pripomore predvsem k boljšemu razumevanju vpliva pojasnjevalnih spremenljivk na škodno pogostost ter povprečno škodo.

Osredotočimo se na modeliranje zavarovanja avtomobilske odgovornosti (AO), ki je v Sloveniji obvezno za vse imetnike avtomobila.

4.1 Analiza podatkov

4.1.1 Predanaliza

Po mnenju avtorjev v [1] je kredibilnost rezultatov dosežena s 100.000 podatki. Prav tako je pomembno, da so podatki zbrani za dve ali več let, saj bi podatki, ki se nanašajo zgolj na eno leto lahko bili pod vplivom dogodkov v tistem letu. Za naše modeliranje so podatki pridobljeni s strani ene izmed večjih zavarovalnic v Sloveniji in sicer za 200.000 zavarovalnih polic, sklenjenih v letih 2017–2019. S tem pa smo zadostili zgoraj naštetim pogojem.

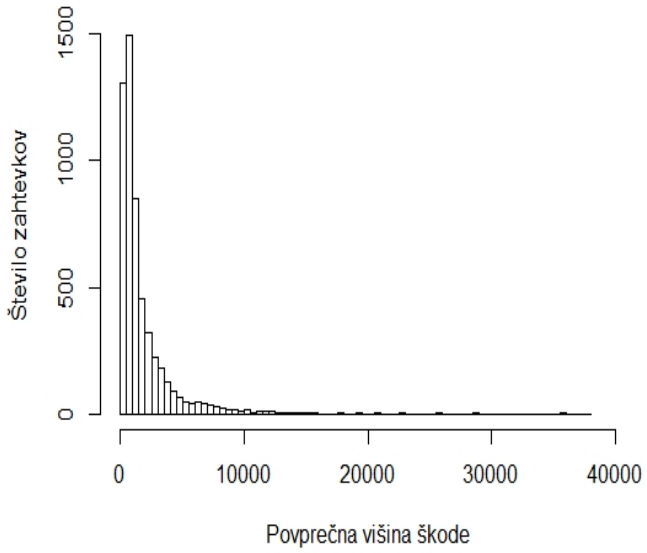
V podatkovni datoteki imamo poleg podatkov o višini škode in število škod za AO, vrednosti za 9 neodvisnih spremenljivk (razred bonus-malus, dejansko število dni kritija, poštno

številko, leto rojstva zavarovanca, prodajni kanal, vrsta plačila, indikator ali je bila polica obnovljena ter letnik in moč vozila). Enota za izpostavljenost je eno leto (365 dni). Izpostavljenost posamezne police je izražena kot delež enega leta. Ločevanje premije glede na spol v Sloveniji od 21. 12. 2012 ni dovoljeno [18], zato podatka o spolu ne vključimo.

Podatke najprej očistimo. Pogledamo ali so vsi podatki smiselni; preverimo zalogo vrednosti posamezne pojasnjevalne spremenljivke (npr. višina premije mora biti pozitivno število). Ugotovimo, da se pojavljajo nepravilnosti pri določenih spremenljivkah, npr. pri letu rojstva in višini premije. Nepravilne vrednosti popravimo oz. izbrišemo in jih obravnavamo kot manjkajoči podatek. Naredimo smiselne modifikacije spremenljivk (npr. leto rojstva spremenimo v starost). Prav tako iz spremenljivke poštna številka generiramo novo spremenljivko *območje*. Za vsako poštno številko izračunamo škodno pogostost (število škod delimo z izpostavljenostjo) in nato združimo tiste poštne številke, ki imajo podobno škodno pogostost in tako dobimo spremenljivko s sedmimi nivoji.

V prečiščeni bazi se je škoda pojavila pri 2,8 % zavarovalnih policah.

S pomočjo grafov preverimo porazdelitev posameznih pojasnjevalnih spremenljivk. Prikažimo graf porazdelitve povprečne višine škode v naši podatkovni bazi. Opazimo izrazito asimetrično porazdelitev v desno, ki je tudi sicer tipična za povprečno višino škode.



Slika 4.1: Porazdelitev povprečne višine škode

4.1.2 Manjkajoči podatki in grupiranje spremenljivk

V podatkovni datoteki, ki jo obravnavamo, se pojavljajo tudi manjkajoči podatki, predvsem pri spremenljivkah leto rojstva, letnik in moč vozila. Z brisanjem vrstic bi izgubili kar velik delež podatkov, zato se raje poslužimo drugih metod.

Uporabimo lahko metodo enkratnega vstavljanja, kjer manjkajočo vrednost nadomestimo npr. s povprečjem, z modusu oz. z mediano. Zavedati pa se moramo, da nadomeščene vrednosti niso nujno dejanske vrednosti. Slabost tega pristopa je ta, da so vstavljene vrednosti pri nadaljnji analizi obravnavane enako kot dejanski podatki in se ne upošteva negotovost uporabljene tehnike. To slabost odpravlja bolj sofisticirana metoda (večkratnega) vstavljanja (angl. *(multiple) imputation*), kjer nadomestimo manjkajoče podatke z napovednimi vrednostmi. Zgradimo model, ki se uči na podmnožici pravilnih (neproblematičnih) podatkov in napove manjkajočo vrednost. V programu R za ta namen uporabimo funkcijo $mice()$, kjer lahko določimo metodo, ki se bo izvedla nad posameznim tipom spremenljivke. S to funkcijo lahko generiramo več baz, na vseh bazah uporabimo posplošen linearni model, nato pa povprečimo rezultat.

Drugi pristop, ki je uporabljen tudi na naši podatkovni bazi, je združitev manjkajočih podatkov v en razred znotraj spremenljivke. Nato izvedemo funkcijo $glm()$ in priključimo manjkajoče spremenljivke h kategoriji, ki ima podoben regresijski koeficient.

Modeliranje z GLM dovoljuje uporabo tako numeričnih kot kategoričnih spremenljivk. V praksi za numerične spremenljivke generiramo ustrezne kategorične spremenljivke, ki jim določimo smiselne razrede. Pri tem je potrebno paziti, da je v vsakem razredu zadostno število podatkov. Včasih je smiselno zmanjšati tudi število nivojev v kategoričnih spremenljivkah. S tem izgubimo določene informacije, vendar model postane enostavnejši in bolj pregleden. Te modifikacije so narejene v procesu izdelave modela.

4.1.3 Delitev podatkov

Podatke je priporočljivo razdeliti (vsaj) v dve skupini: učno in testno množico. Prva, ki ponavadi predstavlja 60 % ali 70 % vseh podatkov (izbranih naključno ali na podlagi časovne spremenljivke, kot je npr. koledarsko leto), je uporabljena za izvajanje vseh korakov oblikovanja modela; od izbire pojasnjevalnih spremenljivk do izpopolnjevanja modela. Ko je model zgrajen, s pomočjo testne skupine ocenimo uspešnost modela. Odločitev, koliko odstotkov podatkov bomo uporabili v učni oz. testni množici, vključuje kompromis. Več podatkov za učenje bo pripomoglo k boljšemu prepoznavanju vzorcev. Vendar, če ostane premalo podatkov za testiranje, bo končna ocena modela manj zanesljiva.

Alternativa zgornji delitvi podatkov je prečno preverjanje. Ta postopek uporablja večkratno učenje in testiranje. Obstaja več različic, največkrat se uporablja k-kratno prečno preverjanje (angl. *k-fold cross-validation*), katerega postopek je naslednji. Najprej se podatki razdelijo v k disjunktnih podmnožic, kjer je k poljubno določen (pogosta izbira je $k = 10$). Delitev je običajno izvedena naključno. Nato vsako podmnožico uporabimo kot testno skupino, preostalih $k - 1$ podmnožic uporabimo kot učno množico. Rezultat tega postopka je k ocen uspešnosti modela, od katerih je bila vsaka ocenjena na podlagi novih (ne-videnih) podatkov. Te ocene točnosti nato povprečimo v končno oceno.

Za večino napovednih modelov, in pri strojnem učenju, je slednji pristop boljši od delitve na učno in testno množico, saj se vsak primer pojavi tako v učni kot v testni množici. Vendar ima k-kratno prečno preverjanje pri modeliranju zavarovalniških podatkov omejeno uporabnost. Postopek ima pomembno omejitev; da je učinkovit, mora faza treniranja obsegati vse korake oblikovanja modela, tudi proces izbire pojasnjevalnih spremenljivk, kar je glavnina modeliranja pri GLM. Uporaba prečnega preverjanja je primerna, kadar se uporabi avtomatizirana izbira pojasnjevalnih spremenljivk. Pri modeliranju v zavarovalništvu je praksa, da se pojasnjevalne spremenljivke izberejo ročno s skrbno presojo. V [5] se priporoča naj bo prednostni pristop delitev podatkov na učno in testno množico, končni model pa se nato zgradi na celotni množici podatkov [5].

4.2 Modeliranje škodne pogostosti

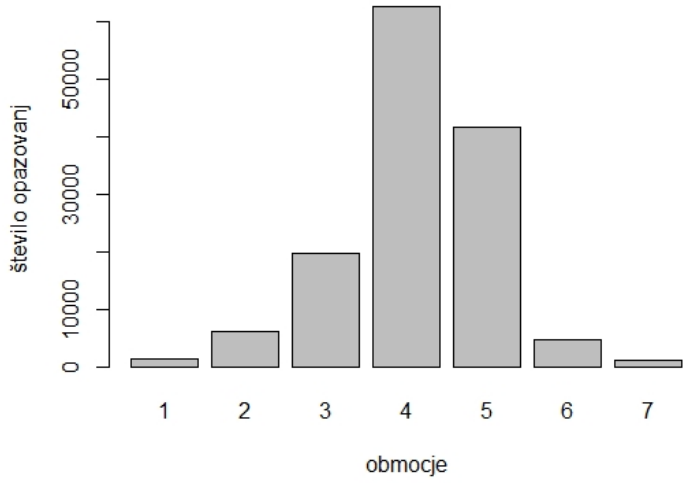
Za modeliranje škodne pogostosti uporabimo najbolj pogosto izbiro porazdelitve in vezne funkcije v literaturi, torej Poissonovo porazdelitev in logaritemsko vezno funkcijo. Odvisna spremenljivka je število škod, za znan efekt pa uporabimo logaritem izpostavljenosti. Model gradimo na učni množici, ki predstavlja 70 % naključno izbranih podatkov.

V model vključimo spremenljivke: moč motorja (*moc_motorja*), starost vozila (*starost_vozila*), BM razred (*BM*), starost zavarovanca (*starost*), indikator ali je polica obnovljena (*obnova*), vrsta plačila (*vrsta_placila*), vrsta zavarovanca (*vrsta_zav*), prodajni kanal (*prod_kanal*) in območje (*obmocje*).

V programu R se za posplošene linearne modele uporabi funkcija v obliki

$$glm(formula; family(link = veznaFunkcija); data = podatki).$$

V teoretičnem delu smo omenili, da je potrebno paziti, da je za bazno kategorijo izbrana kategorija z dovolj podatki. Podkrepimo to na primeru spremenljivke *obmocje*. Porazdelitev te spremenljivke na učni množici je prikazana na grafu 4.2.

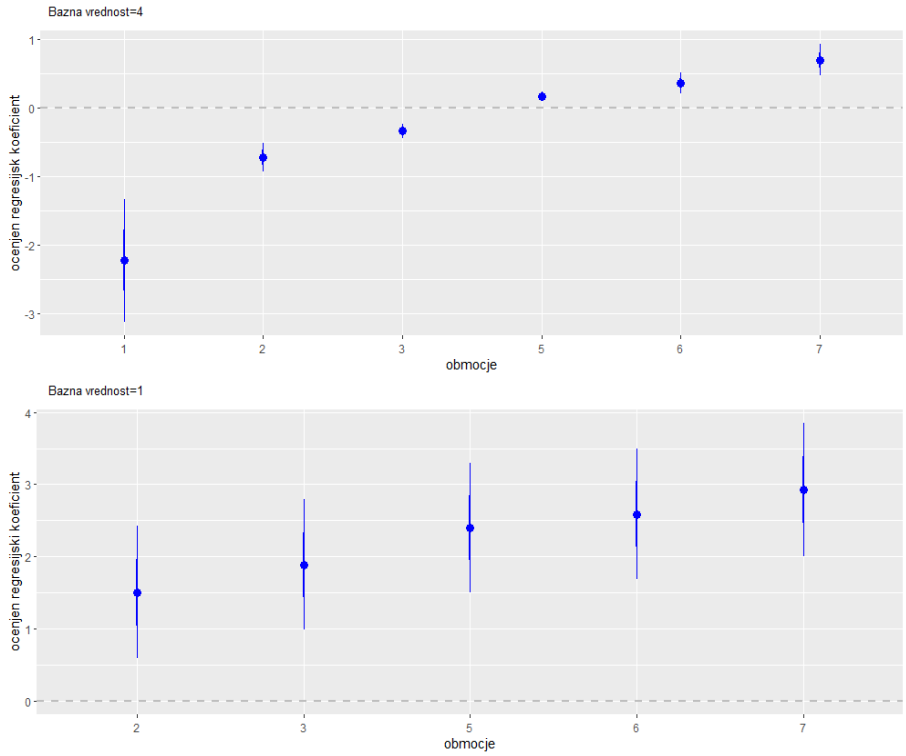


Slika 4.2: Porazdelitev pojasnjevalne spremenljivke *območje*

Najprej za bazno kategorijo nastavimo tisto, ki vsebuje največ opazovanj (kategorija 4). Slika 4.3 prikazuje ocenjene regresijske koeficiente skupaj s 95-odstotnimi intervali zaupanja za vsako kategorijo (bazna kategorija ni prikazana). V programu R je za bazno kategorijo privzeto uporabljena prva kategorija posamezne kategorične pojasnjevalne spremenljivke. Poglejmo, kaj se zgodi, če pri spremenljivki *območje* uporabimo privzeto bazno kategorijo 1, ki ima zelo malo opazovanj v primerjavi z ostalimi (slika 4.2). Vsi ocenjeni regresijski koeficienti so še vedno statistično značilni, vendar je standardna napaka ocen velika; posledično so intervali zaupanja širši in ocenjeni regresijski koeficienti so manj zanesljivi. Prav zaradi tega vsem pojasnjevalnim spremenljivkam določimo za bazno kategorijo tisto, ki ima največji delež opazovanih enot.

Preverimo statistično značilnost vključenih spremenljivk s pomočjo analize deviance in kriterija AIC, kot je bilo opisano v teoretičnem delu naloge. Primerjamo gnezdena modela; model z vsemi spremenljivkami z modelom, kjer smo izpustili eno pojasnjevalno spremenljivko. Zanima nas, ali dodatna pojasnjevalna spremenljivka bistveno prispeva k pojasnitvi modela. V programu R za ta namen uporabimo funkcijo *drop*1(), ki ji kot argument določimo χ^2 test.

Iz slike 4.4 je razvidno, da je vrednost AIC modela z vsemi pojasnjevalnimi spremenljivkami enaka 37.113. Vrednost AIC za model brez spremenljivke *prod_kanal* se bi znižala na 37.099, kar pomeni, da bi dobili ustreznejši model. Opazimo, da ta spremenljivka ni statistično značilna pri stopnji značilnosti $\alpha = 0,05$. A to ne pomeni nujno, da nima vpliva na odvisno spremenljivko. Spremenljivka je lahko označena za nesignifikantno tudi zaradi



Slika 4.3: Ocenjeni regresijski koeficienti in 95-odstotni intervali zaupanja za posamezno območje

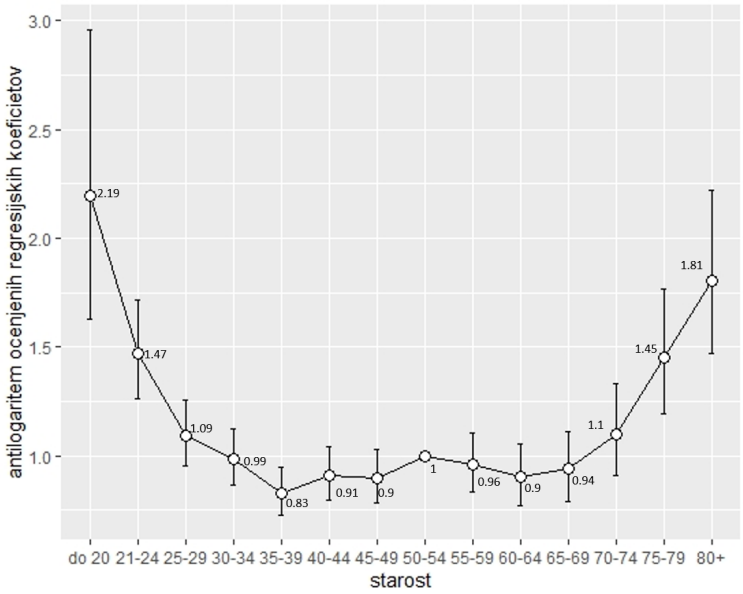
```
Single term deletions

Model:
st_spisov_ao ~ obnova + vrsta_placila + obmocje + starost + moc_motorja +
    starost_vozila + BM + vrsta_zav + prod_kanal
              Df Deviance   AIC     LRT  Pr(>Chi)
<none>             28617 37113
obnova         1    28628 37123  11.796 0.0005936 ***
vrsta_placila  1    28664 37159  47.901 4.484e-12 ***
obmocje        6    28887 37371 269.992 < 2.2e-16 ***
starost       14    28755 37224 138.740 < 2.2e-16 ***
moc_motorja   14    28656 37125  39.612 0.0002932 ***
starost_vozila 26    28683 37127  65.979 2.507e-05 ***
BM            11    28701 37175  84.190 2.271e-13 ***
vrsta_zav      1    28621 37115   4.245 0.0393585 *
prod_kanal    14    28630 37099  13.727 0.4702187
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Slika 4.4: Izpis rezultatov funkcije *drop*1()

prevelikega števila kategorij. Preden jo popolnoma izključimo iz modela, preverimo ali bi bilo smiselno združiti posamezne kategorije. Podobno kot prej, izračunamo intervale zaupa-

nja in jih skupaj z regresijskimi koeficienti prikažemo na grafu. Izkaže se, da pri nekaterih kategorijah dobimo zelo široke intervale zaupanja, kar je posledica malo opazovanih enot v tisti kategoriji. V kolikor interval zaupanja ene kategorije vsebuje interval zaupanja druge kategorije, kategoriji združimo. Z združevanjem kategorij kot končen rezultat dobimo dve kategoriji in s tem tudi statistično značilno spremenljivko. Zaključimo, da pojasnjevalne spremenljivke *prod_kanal* ne izločimo iz modela.



Slika 4.5: Relativnosti za spremenljivko *starost*

Pri modeliranju škodne pogostosti nas bolj kot regresijski koeficienti β zanimajo relativnosti γ. Te dobimo tako, da antilogaritmiramo regresijske koeficiente; $\gamma_j = e^{\beta_j}$. Pri kategoričnih spremenljivkah je relativnost bazne kategorije enaka 1.

Poglejmo si razlago vseh relativnosti na primeru spremenljivke *starost*, ki so prikazane na grafu 4.5. Na podlagi vzorčnih podatkov ocenjujemo, da je regresijski koeficient za kategorijo 80+ enak $0,592$. To pomeni, če vse ostale pojasnjevalne spremenljivke ostanejo nespremenjene, je ocenjena relativnost za to kategorijo $81\,\%$ večja ($e^{0,592} = 1,81$), kot je pričakovano za bazno kategorijo (50–54). Povedano drugače, relativnost $\gamma_j = e^{\beta_j}$ je dodatek za kategorijo j.

Iz zgornjega grafa je razvidno tudi, da imajo najvišjo škodno pogostost mladi, stari do dvajset let, nato sledijo starejši zavarovanci (nad 80 let). Opazimo lahko, da ima starostna skupina 50–54 večjo škodno pogostost kot zavarovanci v sosednjih starostnih skupinah. Temu gre morda pripisati razlog, da so v tej starostni skupini velikokrat starši (katerih otroci so že polnoletni), ki posodijo avtomobil svojim otrokom. Ti zaradi malo izkušenj z vožnjo spadajo v rizično skupino.

Relativnosti izračunamo za vsako pojasnjevalno spremenljivko in jih prikažemo v tabeli. Za ilustracijo prikažimo zgolj relativnosti za starost in območje (tabela 4.1). Relativnost škodne pogostosti za posameznika določimo tako, da množimo relativnosti vsake spremenljivke iz ustreznega razreda. Npr. osebi, ki je stara 35 let in živi v območju 3, pripišemo relativnost $0,83 \cdot 0,71 = 0,59$.

Spremenljivka	**Razred**	**Ocenjena relativnost**
starost	do 20	2,19
	21–24	1,47
	25–29	1,09
	30–34	0,99
	35–38	0,83
	40–44	0,91
	45–49	0,90
	50–54	1
	55–59	0,96
	60–64	0,9
	65–69	0,94
	70–74	1,10
	75–79	1,45
	80+	1,81
obmocje	1	0,11
	2	0,48
	3	0,71
	4	1
	5	1,19
	6	1,43
	7	2

Tabela 4.1: Ocenjene relativnosti pri spremenljivkah *starost* in *obmocje*

Preverimo, če je v modelu prisotna prerazpršenost oz. podrazpršenost. Matematično upanje odvisne spremenljivke je enako 0,04, zato ne moremo uporabiti pravila zidarskega palca o razmerju med devianco in prostostnimi stopnjami. Preverimo s formalnim testom; v programu R uporabimo funkcijo *dispersiontest*(). Ideja testa je razložena v poglavju 3.11. Ničelno hipotezo, da je $c = 0$, na podlagi tega testa (slika 4.6), pri stopnji značilnosti 0,05 zavrnemo, in sprejmemo sklep, da je $c > 0$. V našem modelu je torej prisotna prerazpršenost.

Odločimo se za uporabo kvazipoissonovega modela. Pri izpisu rezultatov se sedaj izpiše tudi vrednost za disperzijski parameter, ki znaša 1,1. To pomeni, da je potrebno standardne napake regresijskih koeficientov v tem modelu pomnožiti s $\sqrt{1,1} = 1,05$.

```
        Overdispersion test

data:  model
z = 2.7099, p-value = 0.003365
alternative hypothesis: true alpha is greater than 0
sample estimates:
     alpha
0.01389118
```

Slika 4.6: Rezultati testa za prerazpšenost

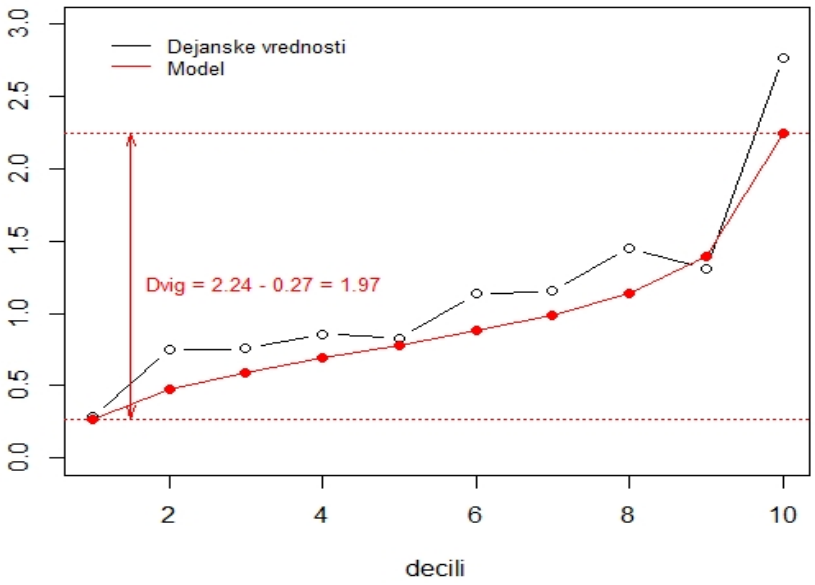
4.2.1 Preverjanje ustreznosti modela

Zgrajen model sedaj testiramo na testni množici. Za vsako opazovano enoto napovemo škodno pogostost in nato uredimo bazo glede na napovedne vrednosti (od najmanjše do največje vrednosti). Urejeno bazo grupiramo v deset skupin, tako da je v vsaki skupini enaka izpostavljenost. Za vsako skupino izračunamo povprečno dejansko in povprečno napovedno vrednost. Nato s pomočjo grafa primerjamo ti dve vrednosti znotraj posamezne skupine. Na sliki 4.7 abcisna os predstavlja vsako skupino, ordinatna os pa predstavlja normalizirane povprečne (dejanske/napovedne) vrednosti (vse vrednosti delimo s skupno povprečno napovedno vrednostjo) [22].

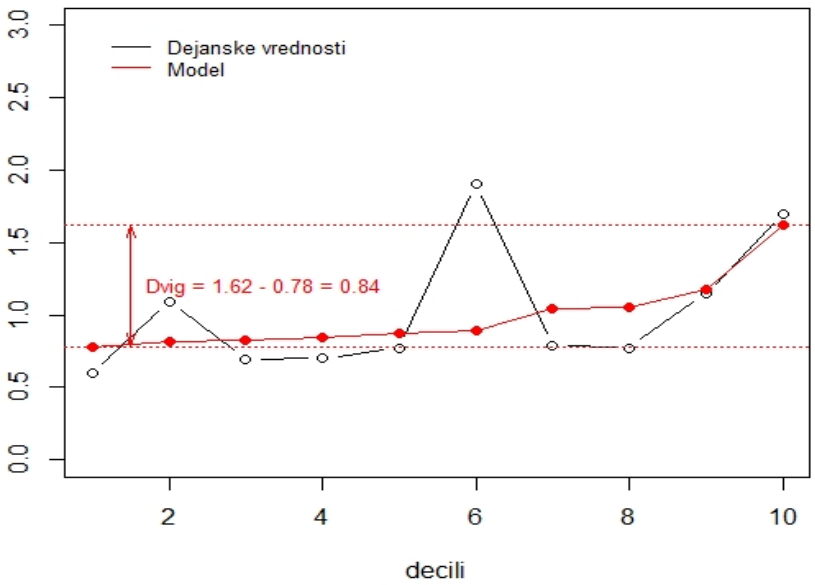
Z večanjem vrednosti skupine se povečujejo povprečne napovedne vrednosti, saj smo podatke uredili v skupine glede na napovedne vrednosti. Če je naš model dober, se bodo povečevale tudi dejanske vrednosti in bodo enakih vrednosti kot napovedne (nekaj odstopanja je dovoljenega). Iz grafa lahko sklepamo, da je model ustrezen, saj so dejanske vrednosti blizu napovednih vrednosti in ni prisotnih prevelikih odstopanj.

Opazimo, da povprečna napovedna vrednost za prvo skupino znaša $0,27$ skupne povprečne napovedne vrednosti, za deseto skupino pa je povprečna napovedna vrednost $2,24$-kratnik skupne povprečne napovedne vrednosti. Dvig za model (razlika med prvo in zadnjo skupino) je enak $2,24 - 0,27 = 1,97$, kar pomeni, da je razlika med tema skupinama približno $197\,\%$.

Za primerjavo si poglejmo krivuljo dviga za model (slika 4.8), kjer smo vključili le eno pojasnjevalno spremenljivko (*starost*). Opazimo velika odstopanja, predvsem v skupini 6. Dvig za model je sedaj enak $0,84$. To pomeni, da ta model slabše ločuje rizične skupine med seboj; model ima manjšo moč segmentacije. Zaključimo, da model iz slike 4.8 ni primeren za modeliranje škodne pogostosti.



Slika 4.7: Krivulja dviga za izbran model škodne pogostosti.

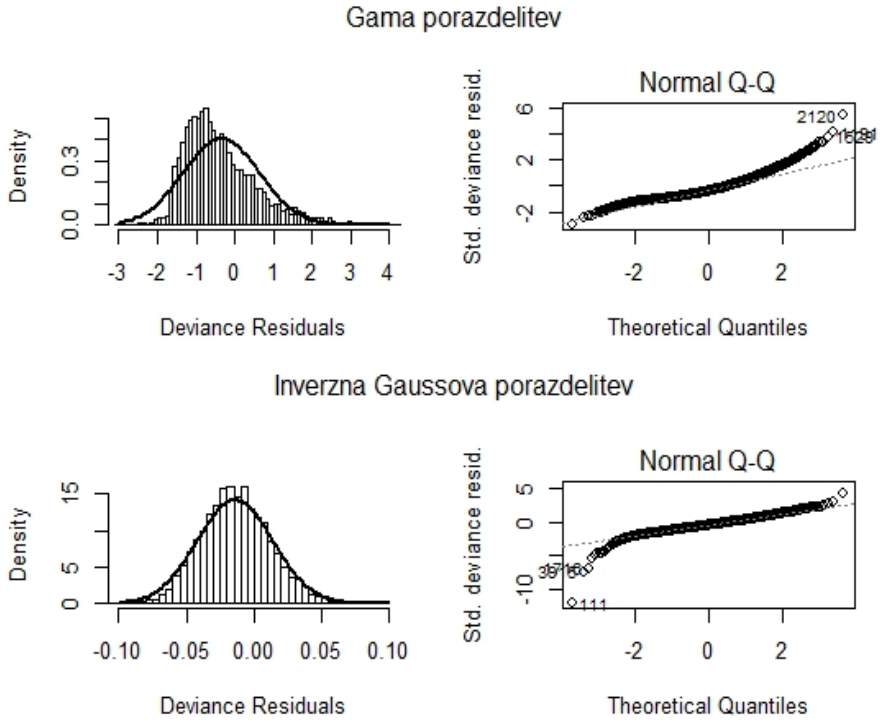


Slika 4.8: Krivulja dviga za model s pojasnjevalno spremenljivko *starost*

4.3 Modeliranje povprečne škode

Za modeliranje povprečne škode uporabimo gama porazdelitev in logaritemsko vezno funkcijo. Posplošen linearen model izvedemo nad podatki, kjer je bila vsaj ena škoda. V model vključimo spremenljivke: moč motorja, vrsta zavarovanca, starost vozila, BM razred in območje. Za utež uporabimo število škod. S postopnim izključevanjem nesignifikantnih pojasnjevalnih spremenljivk dobimo model z zgolj dvema pojasnjevalnima spremenljivkama (moč motorja in BM razred). Ta postopek je analogen kot pri modeliranju škodne pogostosti, le da sedaj pri funkciji *drop*1() uporabimo F-test.

Sedaj na isti množici podatkov uporabimo inverzno Gaussovo porazdelitev z logaritemsko vezno funkcijo. Izkaže se, da kot signifikantne spremenljivke dobimo iste spremenljivke kot pri gama porazdelitvi.



Slika 4.9: Histograma in Q-Q grafikona za gama porazdelitev in inverzno Gaussovo porazdelitev

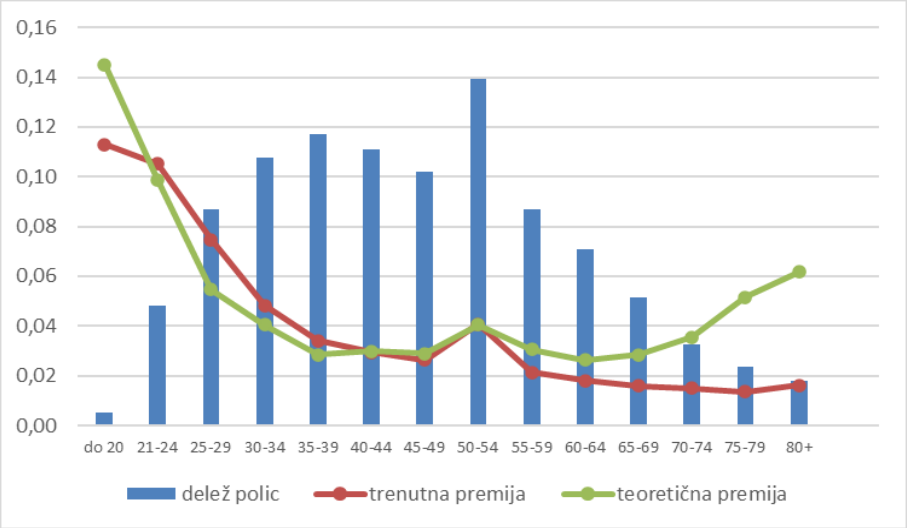
S pomočjo ostankov deviance preverimo, katera porazdelitev odvisne spremenljivke je ustreznejša. Slika 4.9 prikazuje dva primera, kako lahko ocenimo normalnost ostankov deviance; s pomočjo histograma in grafikona Q-Q. Leva stran prikazuje histogram ostankov deviance skupaj z normalno krivuljo, ki se najbolje prilega dani porazdelitvi ostankov. Če izbrana

porazdelitev resnično opiše dane podatke, potem se histogram in krivulja ujemata. Ujemanje opazimo pri inverzno Gaussovi porazdelitvi, medtem ko je v primeru gama porazdelitve histogram asimetričen v desno glede na normalno krivuljo.

Drugi način primerjave porazdelitve ostankov deviance z normalno porazdelitvijo je grafikon Q-Q, ki je prikazan na desni strani slike 4.9. Abcisna os prikazuje teoretične kvantile, ordinatna os pa standardizirane ostanke deviance vzorca. Če so ostanki deviance resnično normalno porazdeljeni, točke opisujejo premico. Da je inverzna Gaussova porazdelitev primernejša izbira za dane podatke, potrjuje tudi grafikon Q-Q, saj se točke bolj prilegajo premici kot pri gama porazdelitvi.

4.4 Primerjava premij

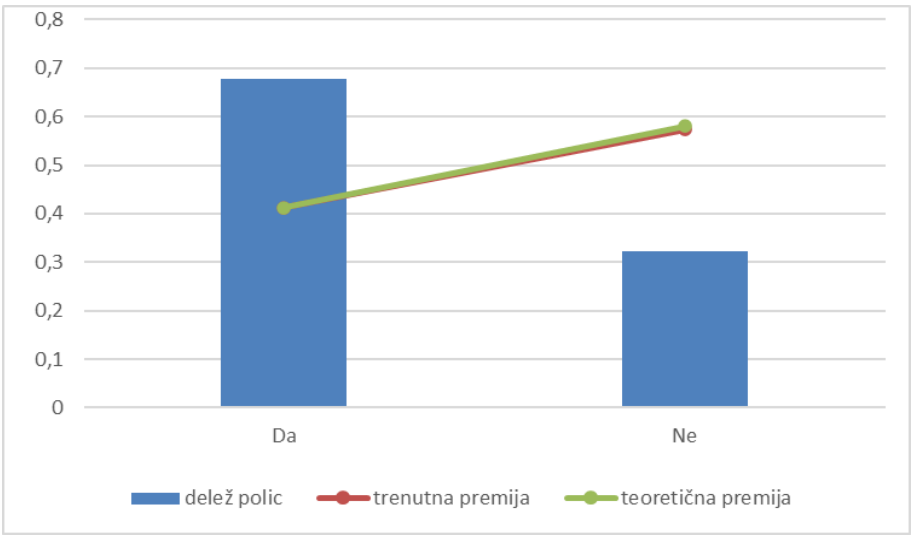
Na celotni množici podatkov uporabimo oba izbrana modela; kvazipoissonov model za škodno pogostost in inverzno Gaussovo regresijo za višino škode. Ta dva rezultata pomnožimo in dobimo teoretično neto premijo. Premijo, pridobljeno s pomočjo GLM modela primerjamo s premijo, ki jo zaračunava izbrana zavarovalnica. Za vsako spremenljivko izračunamo povprečno vrednost trenutne in povprečno vrednost teoretične premije ter primerjamo ti dve vrednosti znotraj posamezne kategorije s pomočjo grafov.



Slika 4.10: Primerjava premij glede na spremenljivko *starost*

Slika 4.10 prikazuje graf primerjave premij glede na spremenljivko *starost*. Rdeča krivulja predstavlja trenutno premijo, zelena krivulja predstavlja premijo pridobljeno s pomočjo posplošenega linearnega modela. Prikazani so tudi deleži polic v portfelju zavarovalnice.

Vrednosti za premijo na grafu ne prikazujemo zaradi varovanja podatkov zavarovalnice. Opazimo, da se v nekaterih kategorijah premiji ujemata (starost od 40 do 54 let), medtem ko bi bilo potrebno za nekatere kategorije obstoječo premijo spremeniti. Najmlajšim in starejšim zavarovancem bi bilo glede na njihovo rizičnost potrebno obstoječo premijo zvišati. Po drugi strani bi zavarovancem med 21 in 39 letom lahko ponudili ugodnejšo premijo. S tem bi najverjetneje privabili več ugodnejših zavarovancev in izgubili zavarovance, ki imajo večjo rizičnost (starost do 20 let) in predstavljajo manjši delež v celotnem portfelju.



Slika 4.11: Primerjava premij glede na spremenljivko *obnova*[1] .

Poglejmo si še primerjavo premij glede na spremenljivko *obnova* (slika 4.11). V tem primeru je razvidno, da zavarovalnica že zaračunava teoretično ustrezno premijo za ta segment.

Na zgoraj opisan način primerjamo premije še za vse ostale pojasnjevalne spremenljivke in ugotovimo, da bi bile modifikacije potrebne za vse preostale segmente.

[1]Vrednost *Da* pomeni, da je stranka predhodno leto že imela sklenjeno zavarovanje z istim avtomobilom. Če zamenja avtomobil, se to obravnava kot novosklenjeno zavarovanje.

Poglavje 5

Uporaba GLM pri analizi prekinitev življenjskih zavarovanj

V drugem delu uporabe posplošenih linearnih modelov na zavarovalniških podatkih analiziramo in poskušamo napovedati kateri zavarovanci bodo predčasno prekinili sklenjeno življenjsko zavarovanje. Zavarovanec ima na voljo več oblik prekinitev (opisano v poglavju 2.2.1). Pri modeliranju med posameznimi oblikami ne bomo ločevali. Pod besedo (predčasna) prekinitev bomo razumeli, da je zavarovanec storniral riziko življenjsko zavarovanje ali pa je pri naložbenem življenjskem zavarovanju izbral eno od naslednjih možnosti; storno ali odkup.

5.1 Podatki in simulacija

Podatki za analizo prekinitev življenjskih zavarovanj so pridobljeni s strani izbrane zavarovalnice. Podatke očistimo, podobno kot smo storili pri podatkih za avtomobilsko zavarovanje. Prečiščena podatkovna baza vsebuje 47.975 zavarovalnih polic, ki so bile sklenjene med letoma 2012 in 2015 ter vsebuje naslednje neodvisne spremenljivke: vrsta zavarovanja (naložbeno ali riziko življenjsko zavarovanje), spol, pristopna starost, višina premije, zavarovalna vsota, frekvenca plačevanja (mesečno, četrtletno, polletno, letno) in prodajni kanal (agencija, ostalo).

Podatke za odvisno spremenljivko (ali je prišlo do predčasne prekinitve) nismo mogli pridobiti, zato si pomagamo z literaturo. Odvisno spremenljivko generiramo tako, da zavarovanci prekinejo zavarovalno polico glede na verjetnost, ki je pridobljena iz regresijskih koeficientov GLM modela iz literature [11]. Število prekinjenih polic tako ne odraža realnega stanja zavarovalnice. S simulacijo ustvarimo enega izmed možnih scenarijev.

Označimo z Y odvisno spremenljivko (*prekinitev*), ki predstavlja odločitev o predčasni prekinitvi zavarovalne police:

$$Y = \begin{cases} 1; & \text{polica predčasno prekinjena} \\ 0; & \text{sicer} \end{cases} . \tag{5.1}$$

Velja, da je slučajna spremenljivka Y porazdeljena binomsko ($Y_i \sim B(1, p_i)$), kjer je p_i verjetnost, da je i-ta polica predčasno prekinjena.

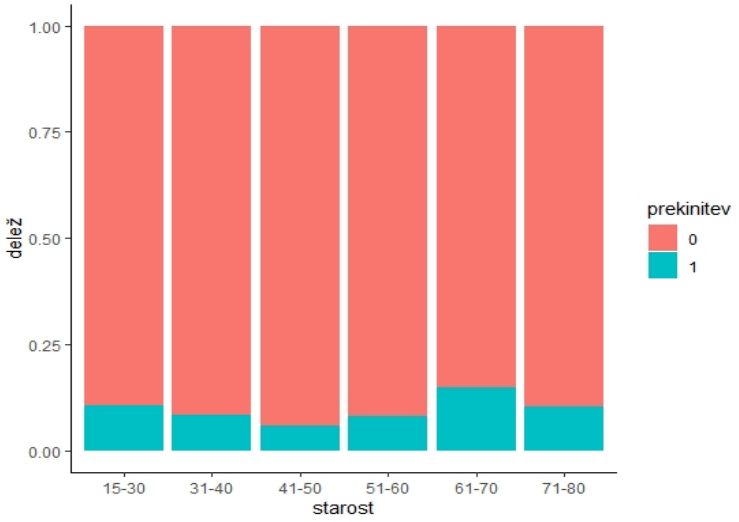
S pomočjo danih regresijskih koeficientov (pridobljenih iz literature [11]) najprej izračunamo linearni prediktor ($X\beta$), nato verjetnost prekinitve in na koncu posamezni vrstici, ki predstavlja zavarovalno polico, pripišemo status z uporabo funkcije $rbinom()$. Ta funkcija simulira izide Bernoullijevih poskusov. Funkcija ima tri argumente: število opazovanj n, število poskusov na opazovanje in verjetnost uspeha za vsako opazovanje.

V programu R izvedemo naslednje ukaze:

$$linpred = X\beta,$$

$$pi = exp(linpred)/(1 + exp(linpred)),$$

$$prekinitev = rbinom(n = n, size = 1, prob = pi).$$



Slika 5.1: Odstotek prekinitev glede na starost

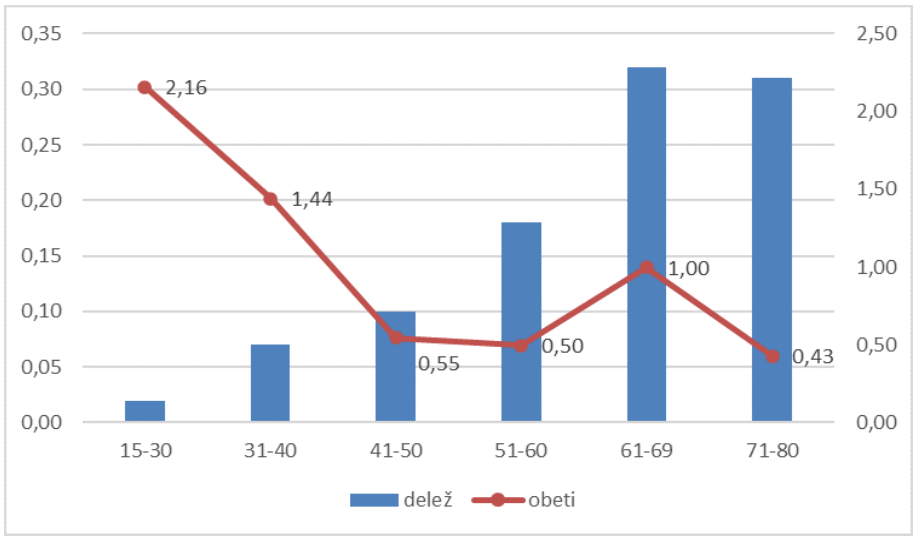
V ustvarjeni bazi je odstotek vseh prekinitev enak 11 %. Napravimo osnovno statistično analizo in pogledamo, kolikšen je delež prekinitev glede na posamezno spremenljivko. Za

zgled prikažimo graf na sliki 5.1, ki prikazuje odstotek zavarovancev, ki so zavarovanje prekinili glede na pristopno starost zavarovanca. Opazimo, da odstotek predčasno prekinjenih polic do starosti 41–50 upada, nato pa se poveča in doseže vrh pri starosti 61–70.

5.2 Grajenje modela

Najprej podatke razdelimo na učno (70 % naključno izbranih podatkov) in testno množico. V model vključimo vse podane pojasnjevalne spremenljivke in izvedemo logistično regresijo na učni množici. Uporabimo binomsko porazdelitev in njeno naravno vezno funkcijo, logit vezno funkcijo. Za bazne kategorije določimo kategorije z največjo zastopanostjo v podatkih. Postopek združevanje posameznih kategorij je analogen kot pri modeliranju škodne pogostosti. Vse pojasnjevalne spremenljivke so se izkazale za statistično značilne pri stopnji značilnosti $\alpha = 0,05$.

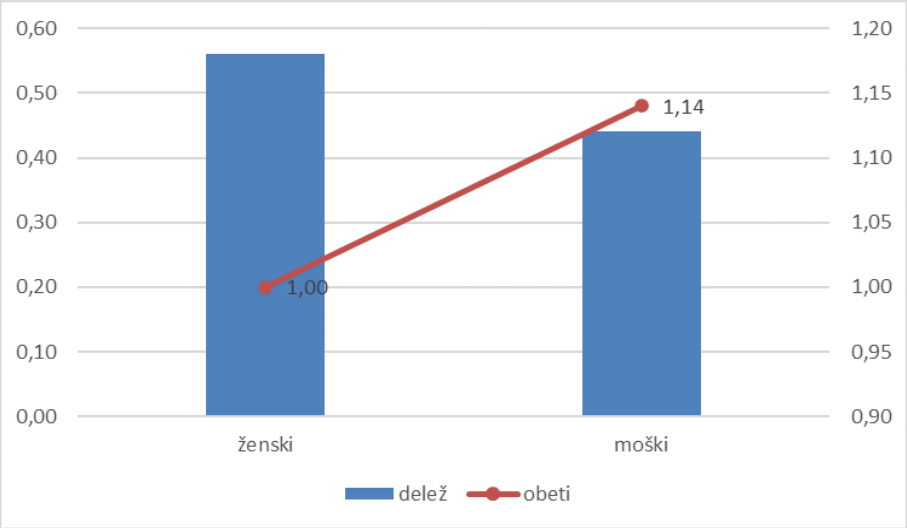
Poglejmo si rezultate modela. Za vsako spremenljivko oblikujemo histogram deležev posameznih kategorij pojasnjevalne spremenljivke skupaj s krivuljo, ki prikazuje obete. Leva ordinatna os ponazarja deleže, medtem ko desna ordinatna os predstavlja vrednost obetov. Grafi torej predstavljajo vpliv pojasnjevalne spremenljivke na odvisno spremenljivko, kjer so upoštevane vse ostale spremenljivke, ki so vključene v model.



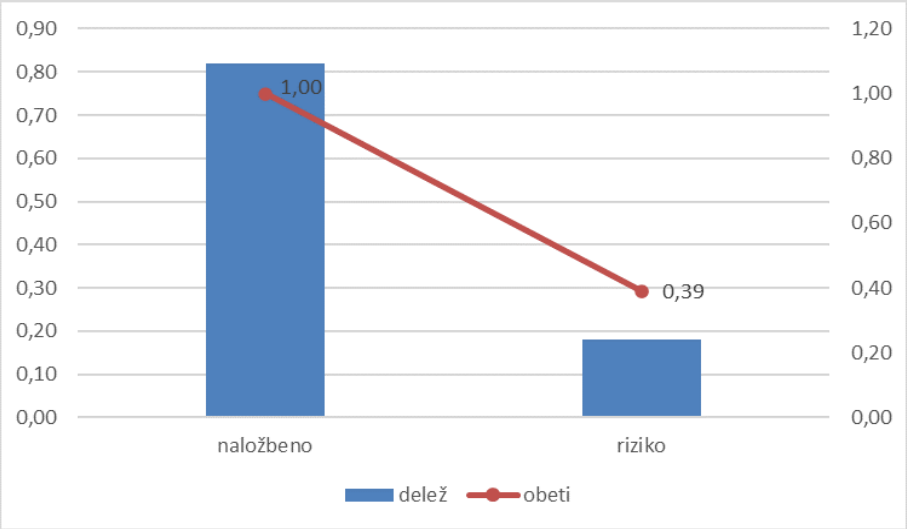
Slika 5.2: Učinek spremenljivke *starost* na prekinitev

Prikažimo grafe za nekatere pojasnjevalne spremenljivke. Na sliki 5.2 je vidno, da je v podatkovni bazi najmanjši delež zavarovancev starih med 15 in 30 let. Obeti za to kategorijo so kar za 116 % večji kot za bazno kategorijo (starost med 61 in 69 let). Obeti za zavarovance stare med 31 in 40 let so večji za 44 % v primerjavi z bazno kategorijo.

Iz slike 5.3 je razvidno, da je obet prekinitve pri moških za 14 % višji kot pri ženskah. Ugotovimo (slika 5.4), da je večja verjetnost prekinitve naložbenega življenjskega zavarovanja. To se zdi smiselno, saj so prihranki zavarovanca vezani na delovanje skladov. Slaba donosnost skladov pa povzroči več predčasnih prekinitev.



Slika 5.3: Učinek spremenljivke *spol* na prekinitev



Slika 5.4: Učinek spremenljivke *produkt* na prekinitev

5.3 Ocena napovedne moči modela

Vsebina podpoglavja je povzeta po literaturi [5].

Logistični model napoveduje, kolikšna je verjetnost, da se bo dogodek zgodil. Vendar je velikokrat v praksi potrebno verjetnost pretvoriti v binarno napoved: ali se bo dogodek zgodil ali ne. V našem primeru želimo določiti ali bo zavarovanec predčasno prekinil zavarovanje ali ne. Za vsakega zavarovanca model vrne verjetnost, s katero bo prekinil zavarovanje. Na podlagi te verjetnosti ga lahko klasificiramo v enega izmed razredov. Pri tem moramo določiti mejno vrednost, nad katero bodo zavarovanci klasificirani kot zavarovanci, ki bodo prekinili zavarovanje.

Uspešnost modela preverimo na testni množici podatkov. Zanima nas, v koliko primerih je model pravilno napovedal izid. Rezultat predstavimo z matriko zamenjav (angl. *confusion matrix*). Vsaka vrstica v matriki zamenjav predstavlja dejanske vrednosti, vsak stolpec pa napovedane vrednosti.

Označimo s FN število lažno negativnih primerov (primer smo napovedali kot negativen, a je v resnici pozitiven), s TN število resnično negativnih primerov (primer smo napovedali kot negativen in je v resnici negativen), s TP število resnično pozitivnih primerov (primer smo napovedali kot pozitiven in je v resnici pozitiven) in s FP število lažno pozitivnih primerov (primer smo napovedali kot pozitiven, a je v resnici negativen).

Prikažimo matriko zamenjav našega modela pri mejni vrednosti 50 %. To pomeni, če velja

$$P(Y = 1|X) > 0,5,$$

potem je $Y = 1$ (pozitiven izid), drugače pa je $Y = 0$ (negativen izid).

	napovedan negativni izid	napovedan pozitivni izid
dejansko negativni izid	$TN = 13275$	$FP = 9$
dejansko pozitivni izid	$FN = 1039$	$TP = 70$

Tabela 5.1: Matrika zamenjav pri mejni vrednosti 50 %

Stopnja natančnosti modela (angl. *accuracy*) je razmerje med pravilnimi napovedmi in skupnim številom primerov in jo izračunamo kot

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}.$$

Če je zgornja vrednost enaka 1, potem model pravilno napove vse izide.

V primeru, da imamo neuravnotežene podatke – to pomeni, da obstaja prevladujoči razred klasifikacije (npr. v naši matriki zamenjav prevladuje resnično negativnih primerov TN) – se raje poslužujemo mer, ki upoštevajo neuravnoteženost. Pogosto se uporabljajo mere natančnost, priklic, F_1-mera ter ROC krivulja.

Natančnost (angl. *precision*) je razmerje med resnično pozitivnimi primeri in med vsemi, ki so klasificirani kot pozitivni:

$$precision = \frac{TP}{TP + FP}.$$

Večja kot je natančnost, manjše je število lažno pozitivnih primerov.

Priklic (angl. *recall, sensitivity*) meri delež resnično pozitivnih primerov med vsemi primeri, ki so dejansko pozitivni:

$$recall = \frac{TP}{TP + FN}.$$

Večji kot je priklic, manjše je število pozitivnih primerov, ki so klasificirani kot negativni.

Nemogoče je imeti tako visoko natančnost kot visok priklic. Ali stremimo k večji natančnosti ali k večjemu priklicu, je odvisno od situacije, ki jo proučujemo.

Na podlagi natančnosti in priklica lahko izračunamo F_1-mero, ki je harmonično povprečje le-teh in da večjo težo meri, ki ima manjšo vrednost. Izračuna se kot

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}.$$

Velika vrednost F_1 sugerira veliko vrednost priklica in natančnosti.

$accuracy$	$0,93$
$precision$	$0,89$
$recall$	$0,06$
F_1	$0,12$

Tabela 5.2: Rezultati modela

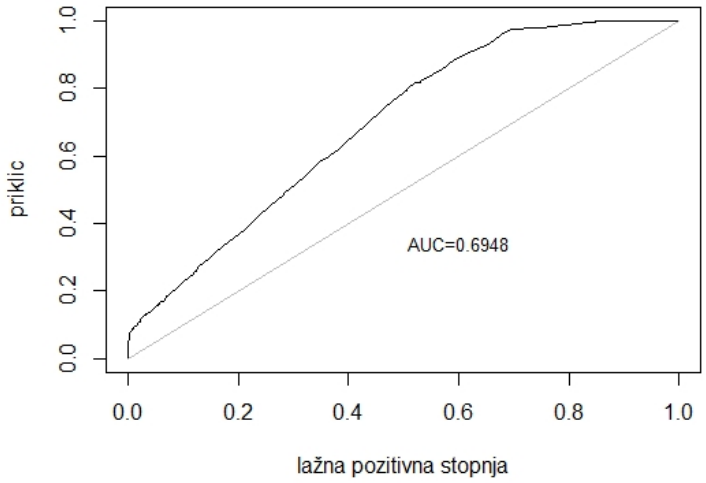
V tabeli 5.2 so prikazane vrednosti nekaterih mer za naš model. Iz tabele je razvidna visoka stopnja natančnosti modela ($accurary$), vendar zaradi neuravnoteženosti ta mera ni ustrezna. Opazimo, da imamo v našem modelu veliko vrednost natančnosti ($precision$) in malo vrednost priklica ($recall$). To pomeni, da je model narobe napovedal veliko dejansko pozitivnih primerov (velik FN), ampak tiste, ki je napovedal kot pozitivne, so resnično pozitivni (majhna vrednost FP).

Lažna pozitivna stopnja (angl. *false positive rate*) meri kako pogosto model napove pozitiven izid, kadar je dejansko negativen:

$$false\ positive\ rate = \frac{FP}{FP + TN}.$$

ROC krivulja (angl. *Receiver Operating Characteristic curve*) je orodje, ki se pogosto uporablja pri binarni klasifikaciji. Prikazuje priklic glede na lažno pozitivno stopnjo.

Mera AUC (angl. *Area under the curve*) je ploščina pod ROC krivuljo. Večja kot je ploščina pod ROC krivuljo, boljši je model. Če je $AUC = 1$, potem imamo opravka s popolnim modelom. V najslabšem primeru je $AUC = 0,5$; takrat model razvršča naključno (model brez napovedne moči).



Slika 5.5: ROC krivulja

Vrednost AUC za naš model je enaka $0,6948$ (slika 5.5), kar pomeni, da model napoveduje bolje kot naključno, vendar ne moremo govoriti o modelu z veliko napovedno močjo.

Za boljše rezultate bi bilo potrebno vključiti še kakšen dejavnik (podatek o tem, po kolikšnem času od sklenjenega zavarovanja se zavarovanci odločijo za predčasno prekinitev, koledarsko leto prekinitve in podobno). Prav gotovo bi bil pomemben tudi podatek o tem ali je zavarovanec zaposlen oz. finančno stanje zavarovanca, vendar so takšni podatki zaupne narave in jih zavarovalnica ne hrani. Smiselno bi bilo tudi raziskati in dodati primerne interakcije med pojasnjevalnimi spremenljivkami, kar pa presega okvirje tega magistrskega dela.

Poglavje 6

Sklep

Danes je upoštevanje vseh pomembnih dejavnikov za izračun premije avtomobilskega zavarovanja ključnega pomena vsake zavarovalnice, saj s tem lažje konkurira na trgu. Premija, ki jo zavarovanec plača zavarovalnici v zameno prevzetja njegovega tveganja, temelji na pričakovani vrednosti njegove škodne pogostosti in višini škode.

V magistrski nalogi smo pokazali, da lahko premijo, ki ustreza stopnji tveganja vsakega posameznika, pridobimo z uporabo posplošenih linearnih modelih na zgodovinskih podatkih zavarovalnice. GLM modeli namreč omogočajo modeliranje podatkov, ki niso nujno normalno porazdeljeni. Pri zavarovalniških podatkih pogosto naletimo na porazdelitve odvisne spremenljivke, ki imajo izrazito asimetrično porazdelitev (npr. gama porazdelitev, inverzna Gaussova porazdelitev) ali na diskretno porazdelitev (npr. binomska porazdelitev, Poissonova porazdelitev). Z ustrezno izbiro porazdelitve odvisne spremenljivke in parametrov modela smo uspešno oblikovali GLM model za določitev poštene premije in s tem dosegli zastavljen cilj. S primerjavo premij smo pokazali, da bi za nekatere zavarovance bilo potrebno premijo povišati, medtem ko bi lahko manj rizičnim zavarovancem ponudili ugodnejšo premijo. S tem bi zavarovalnica privabila več ugodnejših zavarovancev, hkrati pa izgubila nedobičkonosne zavarovance in si izboljšala portfelj.

Uporaba posplošenih linearnih modelov v zavarovalništvu ni omejena samo na oblikovanje premij za neživljenjska zavarovanja. Z njimi lahko analiziramo, kateri zavarovanci so nagnjeni k predčasni prekinitvi življenjskega zavarovanja. V tem primeru uporabimo logistično regresijo in zgradimo klasifikacijski model. Podatka, kateri zavarovanci so predčasno prekinili življenjsko zavarovanje, s strani izbrane zavarovalnice žal nismo mogli pridobiti, zato je zgrajen logistični model temeljil na simulaciji odvisne spremenljivke. S pomočjo modela smo analizirali vplive posameznih spremenljivk na predčasno prekinitev zavarovanja. Vsekakor bi z vključitvijo več pomembnih dejavnikov in dodanimi interakcijami model lahko še izboljšali in mu povečali napovedno moč.

Literatura

[1] D. Anderson, S. Feldblum, C. Modlin, D. Schirmacher, E. Schirmacher, N. Thandi, *A Practitioner's Guide to Generalized Linear Models*, Third edition, A CAS Study Note, 2017.

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

[3] A. J. Dobson, A. G. Barnett, *An Introduction to Generalized Linear Models*, 4. edition, Chapman and Hall/CRC, 2018.

[4] J. Garrido, J. Zhou, *Full Credibility with Generalized Linear and Mixed Models*, 2009, str. 62–80.

[5] M. Goldburd, A. Khare, D. Tevet, *Generalized Linear Models for Insurance Rating*, Casualty Actuarial Society, 2016.

[6] M. Hladnik, *Verjetnost in statistika: zapiski predavanj*, Fakulteta za računalništvo in informatiko, Ljubljana, 2002.

[7] H. Krašovec, *Angleško-slovenski in slovensko-angleški slovar zavarovalništva*, Pegaz International, Ljubljana, 2006.

[8] D. Lillis, *Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output*, pridobljeno: 12. 10. 2019. Dostopno na naslovu: `https://www.theanalysisfactor.com/r-glm-model-fit`.

[9] L. X. Lim, H. M. Kang, *A Comparison of Classification Models for Life Insurance Lapse Risk*, International Journal of Recent Technology and Engineering, Blue Eyes Intelligence Engineering & Sciences Publication, 2019, str. 245–250, pridobljeno: 11. 7. 2019. Dostopno na naslovu: `https://www.researchgate.net/publication/333262288_A_Comparison_of_Classification_Models_for_Life_Insuranceq_Lapse_Risk`.

[10] P. McCullagh, J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, 1989.

[11] X. Milhaud, S. Loisel, V. Maume-Deschamps, *Surrender Triggers in Life Insurance: Classification and Risk Predictions*, 2010. Pridobljeno: 25. 11. 2020. Dostopno na naslovu: `https://www.researchgate.net/publication/41118079_Surrender_Triggers_in_Life_Insurance_Classification_and_Risk_Predictions`.

[12] M. Milič, Finance, Zakaj ljudje bežijo iz življenjskih zavarovanj, pridobljeno: 13. 7. 2019. Dostopno na naslovu: `https://www.finance.si/336283?cctest&`.

[13] P. Močivnik, *Razlaga zavarovalniških izrazov*, Slovensko zavarovalno združenje, 2010, pridobljeno: 11. 7. 2019. Dostopno na naslovu: `https://www.zav-zdruzenje.si/sredisce-informacij`.

[14] E. Ohlsson, B. Johansson, *Non-life Insurance Pricing with Generalized Linear Models*, Springer-Verlag Berlin Heidelberg, 2010.

[15] L. Pfajfar, *Osnovna ekonometrija*, Ekonomska fakulteta v Ljubljani, Ljubljana, 2018.

[16] W. Venables, D. Ripley, *Modern Applied Statistics with S*, 4. edition, Springer Science+Business Media, 2002.

[17] J. Yan, J. Guszcza, M. Flynn, C. Wu, *Applications of the Offset in Property-Casualty Predictive Modeling*, Casualty Actuarial Society E-Forum, 2009, str. 366–385.

[18] PIS, Sklep o podrobnejših pravilih in merilih upoštevanja osebne okoliščine spola, pridobljeno: 12. 10. 2019. Dostopno na naslovu: `http://www.pisrs.si/Pis.web/pregledPredpisa?id=SKLE9487`.

[19] RTV Slovenija, Prekinitev zavarovanja vas lahko drago stane!, pridobljeno: 12. 7. 2019. Dostopno na naslovu: `https://www.rtvslo.si/koda/prekinitev-zavarovanja-vas-lahko-drago-stane/396371`.

[20] Slovensko zavarovalno združenje, Življenjska zavarovanja, pridobljeno: 10. 7. 2019. Dostopno na naslovu: `https://www.zav-zdruzenje.si/skupine-zavarovanj/zivljenjska-zavarovanja-2/`.

[21] Statistični urad Republike Slovenije, Demografsko in socialno področje, pridobljeno: 25. 9. 2019. Dostopno na naslovu: `https://pxweb.stat.si/SiStat`.

[22] Study Note: Model Validation and Holdout Data, The CAS Institute, pridobljeno: 20. 2. 2020. Dostopno na naslovu: `https://thecasinstitute.org/credentials/predictive-analytics-and-data-science/credential-curriculum/`.

[23] Wikipedia, Inverse Gaussian distribution, pridobljeno: 16. 08. 2019. Dostopno na naslovu: `https://en.wikipedia.org/wiki/Inverse_Gaussian_distribution`

[24] Zavarovalnica Sava, Splošni pogoji za življenjsko zavarovanje, pridobljeno: 12. 7. 2019. Dostopno na naslovu: https://www.zav-sava.si/sl-si/zavarovanja/zivljenje/klasicno/.

[25] Zavpro zavarovnje, Naložbeno življenjsko zavarovanje, pridobljeno: 10. 7. 2019. Dostopno na naslovu: http://www.zavpro-zavarovanje.si/nalozbeno_zivljenjsko_zavarovanje.html.